

TEXTE

72/2023

Final report

Review of the OECD Test Guidelines relevant to environmental assessment with regard to the state of the art in science and technology

by:

Dr. Alexandra Polcher
Ramboll Deutschland, Munich

Prof. Dr. Christian Schlechtriem
Fraunhofer IME, Schmallenberg

Dr. Stefan Hahn
Fraunhofer ITEM; Hannover

publisher:

German Environment Agency

TEXTE 72/2023

Ressortforschungsplan of the Federal Ministry for the
Environment, Nature Conservation, Nuclear Safety and
Consumer Protection

Project No. FKZ 3720 64 408 0

Report No. FB001033 ENG

Final report

Review of the OECD Test Guidelines relevant to environmental assessment with regard to the state of the art in science and technology

Klicken Sie hier, um Text einzugeben.

by

Dr. Alexandra Polcher
Ramboll Deutschland, Munich

Prof. Dr. Christian Schlechtriem
Fraunhofer IME, Schmallenberg

Dr. Stefan Hahn
Fraunhofer ITEM; Hannover

On behalf of the German Environment Agency

Imprint

Publisher

Umweltbundesamt
Wörlitzer Platz 1
06844 Dessau-Roßlau
Tel: +49 340-2103-0
Fax: +49 340-2103-2285
buergerservice@uba.de
Internet: www.umweltbundesamt.de

 [/umweltbundesamt.de](https://www.facebook.com/umweltbundesamt.de)

 [/umweltbundesamt](https://twitter.com/umweltbundesamt)

Report performed by:

Ramboll Deutschland GmbH
Werinherstraße 79
81541 München
Germany

Fraunhofer IME
Auf dem Aberg 1
57392 Schmallenberg

Fraunhofer ITEM
Nikolai-Fuchs-Straße 1
30625 Hannover

Report completed in:

August 2022

Edited by:

Section IV 1.1 International Chemicals Management
Susanne Walter-Rohde

Publication as pdf:

<http://www.umweltbundesamt.de/publikationen>

ISSN 1862-4804

Dessau-Roßlau, Mai 2023

The responsibility for the content of this publication lies with the author(s).

Abstract: Review of the OECD Test Guidelines relevant to environmental assessment with regard to the state of the art in science and technology

The OECD Test Guidelines (TG) for Chemicals are a specific tool for assessing the potential effects of chemicals on human health and the environment. These internationally standardised TGs for the testing of chemicals are used by industry, academia and authorities in the testing and evaluation of chemicals (industrial chemicals, pesticides, biocides, pharmaceuticals, etc.). As part of the OECD Test Guidelines Programme (TGP), the OECD TGs are developed and updated by the OECD Working Group of National Coordinators for the OECD Test Guidelines Programme (WNT). To ensure that the OECD TGs reflect the state of the art in science and technology and meet the regulatory requirements of member countries, the OECD TGs are to be continuously expanded and updated. As regular review of the OECD TGs is not a standard requirement, it is the responsibility of OECD member countries to identify, propose and implement the projects necessary to update the TGs. Therefore, the focus is often on guidelines in which member countries have a particular interest and for which sufficient resources are available to initiate a revision process. Test guidelines that may not be used as frequently or are of low interest tend to be neglected in this approach, although a revision may be necessary. The aim of this project was to identify OECD TGs that are not state of the art. Only OECD TGs that refer to the effects of chemicals on biotic systems, to the behaviour and fate of chemicals in the environment or to their physicochemical properties when used for environmental risk assessment were considered. Based on a detailed questionnaire, a comprehensive survey was conducted to identify the needs for TG revisions. Proposals for revising the TGs were evaluated in a second survey. Both surveys were open to all interested parties including industry, science, contract labs, NGOs and authorities. After completion of the surveys, the results were discussed in three thematic workshops. Recommendations for the revision of the OECD Test Guidelines were developed and prioritised, which were finally presented to the members of the WNT.

Kurzbeschreibung: Überprüfung der für die Umweltprüfung relevanten OECD-Prüfrichtlinien im Hinblick auf den Stand von Wissenschaft und Technik

Die OECD-Prüfrichtlinien für die Prüfung von Chemikalien sind ein spezielles Instrument zur Bewertung der möglichen Auswirkungen von Chemikalien auf die menschliche Gesundheit und die Umwelt. Diese international standardisierten Prüfrichtlinien für die Prüfung von Chemikalien werden von der Industrie, der Wissenschaft und den Behörden bei der Prüfung und Bewertung von Chemikalien (Industriechemikalien, Pestizide, Biozide, Arzneimittel usw.) verwendet. Als Teil des OECD-Prüfrichtlinienprogramms (TGP) werden die OECD-Prüfrichtlinien von der OECD-Arbeitsgruppe der nationalen Koordinatoren*Koordinatorinnen für das OECD-Prüfrichtlinienprogramm (WNT) entwickelt und aktualisiert. Um sicherzustellen, dass die OECD-Prüfrichtlinien den Stand von Wissenschaft und Technik widerspiegeln und den regulatorischen Anforderungen der Mitgliedsstaaten entsprechen, sollen die OECD-Prüfrichtlinien kontinuierlich erweitert und aktualisiert werden. Da eine regelmäßige Überprüfung der OECD-Prüfrichtlinien nicht standardmäßig vorgesehen ist, liegt es in der Verantwortung der OECD-Mitgliedsstaaten, die zur Aktualisierung der Prüfrichtlinien notwendigen Projekte zu identifizieren, vorzuschlagen und durchzuführen. Daher liegt der Schwerpunkt häufig auf Richtlinien, an denen die Mitgliedsländer ein besonderes Interesse haben und für die ausreichende Ressourcen zur Verfügung stehen, um einen Überarbeitungsprozess einzuleiten. Prüfrichtlinien, die vielleicht nicht so häufig verwendet werden oder von geringerem Interesse sind, werden bei diesem Ansatz eher vernachlässigt, obwohl eine Überarbeitung erforderlich sein könnte. Ziel dieses Projekts war es die OECD-Prüfrichtlinien zu identifizieren, die nicht dem Stand der Technik entsprechen. Berücksichtigt wurden dabei nur OECD-Prüfrichtlinien, die sich auf die Wirkungen von Chemikalien auf

biotische Systeme, auf das Verhalten und den Verbleib von Chemikalien in der Umwelt oder auf ihre physikalisch-chemischen Eigenschaften beziehen, wenn diese für die Umweltrisikobewertung verwendet werden. Auf der Grundlage eines detaillierten Fragebogens wurde eine umfassende Umfrage durchgeführt, um den notwendigen Revisionsbedarf zu ermitteln. Vorschläge zur Revision der Richtlinien wurden in einer zweiten Umfrage evaluiert. Beide Umfragen standen allen interessierten Kreisen aus Industrie, Wissenschaft, Auftragslaboren, Nicht-Regierungsorganisationen (NGOs) und Behörden offen. Nach Abschluss der Umfragen wurden die Ergebnisse in drei thematischen Workshops diskutiert. Dabei wurden Empfehlungen für die Überarbeitung der OECD-Prüfrichtlinien erarbeitet und priorisiert, die schließlich den Mitgliedern des WNT vorgestellt wurden.

Table of content

List of figures	11
List of tables	11
List of abbreviations	12
Summary	14
Zusammenfassung.....	17
1 Background and objectives	20
2 Project approach	21
3 Data collection.....	22
3.1 Online survey (May 2021 – July 2021)	22
3.1.1 Information needs	22
3.1.2 Conducting the online survey	23
3.2 Pre-evaluation (Nov 2021 – Jan 2022)	25
4 Key results – Suggestions and Pre-evaluation.....	27
4.1 Type of update suggestions	31
4.1.1 Type of suggestions – General	31
4.1.2 Type of suggestions – Editorial	33
4.1.3 Type of suggestions – Scientific specific	34
4.1.4 Type of suggestions – Relevance for test conduction or evaluation	35
4.1.5 Type of suggestions – Relevance for regulation	35
4.2 Pre-evaluation.....	37
4.2.1 Pre-evaluation - rating of relevance of revision	37
4.2.2 Pre-evaluation - feedback on each single update suggestion	37
5 Key results – Relevance and Prioritisation	40
5.1 Block “Fate”	40
5.1.1 Online survey: Number of update suggestions – Top 5 “Fate”	40
5.1.2 Pre-evaluation of suggestions – relevance of TG revision “Fate”	40
5.2 Block “Terrestrial”	44
5.2.1 Online survey: Number of update suggestions – Top 5 “Terrestrial”	44
5.2.2 Pre-evaluation of suggestions – relevance of TG revision “Terrestrial”	44
5.3 Block “Aquatic”	46
5.3.1 Online survey: Number of update suggestions – Top 5 “Aquatic”	46
5.3.2 Pre-evaluation of suggestions – relevance of TG revision “Aquatic”	46
5.4 Relevance of year of publication.....	49

5.5	Interest in OECD Expert Group participation.....	51
5.6	Conclusions on the online and pre-evaluation survey.....	53
5.7	International workshops.....	53
6	The OECD Test Guideline update process.....	57
6.1	Official update process	57
6.2	Ahead of the formal submission of a proposal.....	57
7	List of references.....	59
A	Assessment feedback Block A	63
A.1	Test No. 105 - Water solubility	63
A.2	Test No. 112: Dissociation Constants in Water.....	66
A.3	Test No. 318: Dispersion Stability of Nanomaterials in Simulated Environmental Media ...	67
A.4	Test No. 113: Screening Test for Thermal Stability and Stability in Air	68
A.5	Test No. 107: Partition Coefficient (n-octanol/water): Shake Flask Method	69
A.6	Test No. 117: Partition Coefficient (n-octanol/water), HPLC Method.....	71
A.7	Test No. 123: Partition Coefficient (1-Octanol/Water): Slow-Stirring Method	72
A.8	Test No. 106: Adsorption -- Desorption Using a Batch Equilibrium Method.....	73
A.9	Test No. 121: Estimation of the Adsorption Coefficient (K _{oc}) on Soil and on Sewage Sludge using High Performance Liquid Chromatography (HPLC)	81
A.10	Test No. 312: Leaching in Soil Columns	82
A.11	Test No. 122: Determination of pH, Acidity and Alkalinity	83
A.12	Test No. 115: Surface Tension of Aqueous Solutions	84
A.13	Revised Introduction to the OECD Guidelines for Testing of Chemicals, Section 3.....	84
A.14	Test No. 209: Activated Sludge, Respiration Inhibition Test (Carbon and Ammonium Oxidation).....	90
A.15	Test No. 301: Ready Biodegradability	91
A.16	Test No. 306: Biodegradability in Seawater.....	108
A.17	Test No. 310: Ready Biodegradability - CO ₂ in sealed vessels (Headspace Test).....	109
A.18	Test No. 302B: Inherent Biodegradability: Zahn-Wellens/ EVPA Test.....	111
A.19	Test No. 307: Aerobic and Anaerobic Transformation in Soil.....	113
A.20	Test No. 308: Aerobic and Anaerobic Transformation in Aquatic Sediment Systems.....	125
A.21	Test No. 309: Aerobic Mineralisation in Surface Water – Simulation Biodegradation Test	140
A.22	Test No. 111: Hydrolysis as a Function of pH.....	158
A.23	Test No. 316: Phototransformation of Chemicals in Water – Direct Photolysis.....	159
A.24	OECD Draft Soil Photolysis guideline (2002)	161
A.25	Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure	162

A.26	Test No. 319B: Determination of in vitro intrinsic clearance using rainbow trout liver S9 sub-cellular fraction (RT-S9).....	165
B	Assessment feedback Block B	166
B.1	Test No. 205: Avian Dietary Toxicity Test	166
B.2	Test No. 206: Avian Reproduction Test	168
B.3	Test No. 223: Avian Acute Oral Toxicity Test.....	187
B.4	Test No. 213: Honeybees, Acute Oral Toxicity Test	187
B.5	Test No. 214: Honeybees, Acute Contact Toxicity Test	212
B.6	Test No. 237: Honey Bee (<i>Apis Mellifera</i>) Larval Toxicity Test, Single Exposure	227
B.7	Test No. 245: Honey Bee (<i>Apis Mellifera</i> L.), Chronic Oral Toxicity Test (10-Day Feeding)	230
B.8	Test No. 246: Bumblebee, Acute Contact Toxicity Test.....	233
B.9	Test No. 247: Bumblebee, Acute Oral Toxicity Test.....	236
B.10	Guidance Document 239: Honey Bee Larval	237
B.11	Test No. 208: Terrestrial Plant Test: Seedling Emergence and Seedling Growth Test	243
B.12	Test No. 227: Terrestrial Plant Test: Vegetative Vigour Test.....	258
B.13	Test No. 207: Earthworm, Acute Toxicity Tests	279
B.14	Test No. 216: Soil Microorganisms: Nitrogen Transformation Test	284
B.15	Test No. 220: Enchytraeid Reproduction Test	292
B.16	Test No. 222: Earthworm Reproduction Test (<i>Eisenia fetida</i> / <i>Eisenia andrei</i>)	294
B.17	Test No. 226: Predatory mite (<i>Hypoaspis</i> (<i>Geolaelaps</i>) <i>aculeifer</i>) reproduction test in soil	312
B.18	Test No. 232: Collembolan Reproduction Test in Soil.....	323
B.19	Test No. 228: Determination of Developmental Toxicity to Dipteran Dung Flies(<i>Scathophaga stercoraria</i> L. (<i>Scathophagidae</i>), <i>Musca autumnalis</i> De Geer (<i>Muscidae</i>))	333
C	Assessment feedback Block C	335
C.1	Test No. 203: Fish, Acute Toxicity Test	335
C.2	Test No. 210: Fish, Early-life Stage Toxicity Test.....	345
C.3	Test No. 212: Fish, Short-term Toxicity Test on Embryo and Sac-Fry Stages.....	354
C.4	Test No. 215: Fish, Juvenile Growth Test.....	356
C.5	Test No. 229: Fish Short Term Reproduction Assay.....	362
C.6	Test No. 230: 21-day Fish Assay.....	371
C.7	Test No. 234: Fish Sexual Development Test.....	377
C.8	Test No. 236: Fish Embryo Acute Toxicity (FET) Test	385
C.9	Test No. 240: Medaka Extended One Generation Reproduction Test (MEOGRT)	392

C.10	Test No. 201: Freshwater Alga and Cyanobacteria, Growth Inhibition Test	394
C.11	Test No. 202: Daphnia sp. Acute Immobilisation Test	409
C.12	Test No. 209: Activated Sludge, Respiration Inhibition Test (Carbon and Ammonium Oxidation).....	418
C.13	Test No. 211: Daphnia magna Reproduction Test	420
C.14	Test No. 243: Lymnaea stagnalis Reproduction Test	431
C.15	Test No. 218: Sediment-Water Chironomid Toxicity Using Spiked Sediment	433
C.16	Test No. 219: Sediment-Water Chironomid Toxicity Using Spiked Water.....	437
C.17	Test No. 225: Sediment-Water Lumbriculus Toxicity Test Using Spiked Sediment.....	439
C.18	Test No. 233: Sediment-Water Chironomid Life-Cycle Toxicity Test Using Spiked Water or Spiked Sediment	440
C.19	Test No. 235: Chironomus sp., Acute Immobilisation Test	442
C.20	Test No. 221: Lemna sp. Growth Inhibition Test	443
C.21	Test No. 238: Sediment-Free Myriophyllum Spicatum Toxicity Test.....	452
C.22	Test No. 239: Water-Sediment Myriophyllum Spicatum Toxicity Test.....	456
C.23	Test No. 231: Amphibian Metamorphosis Assay	460
C.24	Test No. 241: The Larval Amphibian Growth and Development Assay (LAGDA).....	471
C.25	Test No. 248: Xenopus Eleutheroembryonic Thyroid Assay (XETA)	472

List of figures

Figure 1:	Geographical distribution of respondents	24
Figure 2:	Distribution for responder groups.....	24
Figure 3:	Survey: Number of returned questionnaires and individual update suggestions “Fate”	28
Figure 4:	Survey: Number of returned questionnaires and individual update suggestions “Terrestrial”	29
Figure 5:	Survey: Number of returned questionnaires and individual update suggestions “Aquatic”	30
Figure 6:	Pre-evaluation of suggestions – Relevance of TG revision (weighted %) vs. Year of publication; Block “Fate”	50
Figure 7:	Pre-evaluation of suggestions – Relevance of TG revision (weighted %) vs. Year of publication; Block “Terrestrial”	50
Figure 8:	Pre-evaluation of suggestions – Relevance of TG revision (weighted %) vs. Year of publication; Block “Aquatic”	51
Figure 9:	Interest in OECD Expert Group participation (Block “Fate”)	52
Figure 10:	Interest in OECD Expert Group participation (Block “Terrestrial”).....	52
Figure 11:	Interest in OECD Expert Group participation (Block “Aquatic”)	53
Figure 12:	OECD Test Guideline Revision Process	57

List of tables

Table 1:	Number of update suggestions	27
Table 2:	Relevance of TG revision (%) in Block “Fate” ranked according to “high relevance”	41
Table 3:	Relevance of TG revision in Block “Fate” ranked according to weighted relevance [%]	42
Table 4:	Relevance of TG revision (%) in Block “Terrestrial” ranked according to “high relevance”	44
Table 5:	Relevance of TG revision in Block “Terrestrial” ranked according to weighted relevance [%]	45
Table 6:	Relevance of TG revision (%) in Block “Aquatic” ranked according to “high relevance”	47
Table 7:	Relevance of TG revision in Block “Aquatic” ranked according to weighted relevance [%].....	48

List of abbreviations

AT	Austria
BE	Belgium
BIAC	Business at OECD
CA	Canada
CBC	OECD Chemicals and Biotechnology Committee
CH	Switzerland
DE	Germany
DK	Denmark
EPOC	Environment Policy Committee
ERA	Environmental risk assessment
EU	European Union
FET	Fish Embryo Acute Toxicity Test
FI	Finland
FR	France
GD	Guidance Document
GLP	Good Laboratory Practice
IT	Italy
JP	Japan
MCs	Member Countries
NCs	National Co-ordinators of the OECD Test Guidelines Programme
NL	Netherlands
NGO	Non-Governmental Organisation
OCSPP	Office of Chemical Safety and Pollution Prevention (US Environmental Protection Agency)
OECD	Organisation for Economic Cooperation and Development
OPPTS	Office of Prevention, Pesticides & Toxic Substances (US Environmental Protection Agency)
REACH	Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals
RBT	ready biodegradability test
SE	Sweden
SETAC	Society of Environmental Toxicology and Chemistry
SI	System International d'unités
TG	Test Guideline
TGP	Test Guideline Programme
UBA	Umweltbundesamt (German Environment Agency)
UK	United Kingdom
US	United States of America
WNT	Working Party of National Co-ordinators of the OECD Test Guidelines Programme
WPB	Working Party on Biocides

WPMN	Working Party on Manufactured Nanomaterials
ZA	South Africa

Summary

OECD Test Guidelines (TGs) display an important tool to evaluate the impacts of chemicals on the environment. The update of the TG is organised by the Working Party of National Co-ordinators for the OECD Test Guidelines Programme (WNT) and the OECD Secretariat, and the procedure relies on proposals from OECD member countries and does not contain a regular update check. To ensure that the OECD TGs reflect the state of the art in science and technology and meet the regulatory requirements of member countries, the OECD TGs are to be continuously expanded and updated. As regular review of the OECD TGs is not a standard requirement, it is the responsibility of OECD member countries to identify, propose and implement the projects necessary to update the TGs. It is therefore not uncommon that only those TGs are updated in which member countries have a particular interest and for which sufficient resources are available to initiate a revision process. It seems obvious that this can lead to neglect of other TGs, although a revision may be necessary.

The German Environment Agency (UBA) contracted Fraunhofer IME, Fraunhofer ITEM and Ramboll to identify and prioritise any potential update needs as regards TGs that refer to the effects of chemicals on biotic systems, to the behaviour and fate of chemicals in the environment or to their physicochemical properties when used for environmental risk assessment. Any potential update needs should be collected based on feedback from the international community.

In a large online survey from May to July 2021, the project team collected 257 completed questionnaires, which included roughly 500 individual update suggestions. In addition, 59 contributions were sent in via email to the project team and were considered further for the project. Originally, it was foreseen that the project team will use developed criteria like number of suggestions received, year of last revision, related time effort, etc. to prioritise the update suggestions. However, due to the high amount of feedback and the very high participation rate, a second survey – the so-called pre-evaluation survey – was conducted in which the international scientific community was confronted with the previously received update suggestions. By giving the international experts the opportunity to see all the suggested update proposals for a particular TG, the project team was then able to ask for a general assessment of the relevance for an update for the TGs. Both surveys were open to all interested parties including industry, science, contract labs, NGOs and authorities.

The relevance of the proposed revisions of the different TGs was assessed by the project team in three different ways. First, the TGs were ranked according to the number of update suggestions made during the online survey. Second, those TGs were identified, which the majority of participants rated high (expressed in %) in the pre-evaluation as regards the relevance of revision. Finally, the rating of the individual TGs (low/medium/high relevance of revision) was further refined by weighting the results obtained. Based on the results obtained by the different ranking approaches, the relevance of revision of the individual guidelines was prioritised.

In a further approach, the relevance of TG revision (weighted %) estimated for each guideline was compared with the TG's year of publication. The results show that no correlation was found between both factors even if only the highest ranked TGs were considered.

Participants of the pre-evaluation survey were asked about their interest in participating in a potential OECD Expert Group dealing with the revision of a specific guideline. Generally, a high interest in OECD Expert Group participation was observed. The results were compared with the relevance of revision estimated for the different TGs. Interestingly, TGs which were identified to have the highest relevance of revision did not necessarily generate the greatest interest in an

OECD Expert Group participation which might be at least partly explained by the limited circle of experts being involved in the usage of specific TGs (e.g., avian toxicity) compared to TGs that are used more generally (e.g., algae or acute fish toxicity).

The project was intended to review which OECD TGs do not correspond to the current state of science and technology. All TGs relevant for the assessment of effects of chemicals on the environment were reviewed. The following conclusions can be drawn:

- ▶ A high number of suggestions to further improve OECD TGs were received (max 55 per TG)
- ▶ A high number of respondents per TG (5-35) contributed to the evaluation of the general relevance of further updating the respective TGs
- ▶ Pre-evaluation of suggestions: Manifold comments which express agreement/disagreement with update suggestions are available and provide thorough basis for further discussion and a potential TG updating process
- ▶ TGs with high relevance of revision were identified in the online pre-evaluation exercise:
 - TOP 3 “Fate and Distribution”
 1. Test No. 301: Ready Biodegradability
 2. Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure
 3. Test No. 310: ready Biodegradability – CO₂ in sealed vessels (Headspace test)
 - TOP 3 “Terrestrial”
 1. Test No. 208: Terrestrial Plant Test: Seedling Emergence and Seedling Growth Test
 2. Test No. 227: Terrestrial Plant Test: Vegetative Vigour Test
 3. Test No. 222: Earthworm Reproduction Test (*Eisenia fetida*/*Eisenia andrei*)
 - TOP 3 “Aquatic”
 1. Test No. 231: Amphibian Metamorphosis Assay
 2. Test No. 201: Freshwater Alga and Cyanobacteria, Growth Inhibition Test
 3. Test No. 234: Fish Sexual Development Test
- ▶ It was shown that the relevance of revision is independent of a TGs’ year of publication
- ▶ There is a high interest in OECD Expert Group participation to further support the further TG revision process

After completion of the surveys, the results were discussed with all interested parties including industry, science, contract labs, NGOs and authorities. Three virtual workshops were conducted covering one of the three different blocks (“fate”, “terrestrial” and “aquatic”) each. Each workshop was offered on two different days and time slots. Important aspects that came out of the survey and pre-evaluation were discussed and enough time for the participants to report on their experiences in conducting the survey was foreseen.

Feedback from the workshops with regard to a future evaluation process for update needs was:

- ▶ Expansion to better involve experts from outside Europe as well as experts from the science and NGOs
- ▶ Extended commenting periods to allow careful revisions of large set of TGs
- ▶ Improved website entry page with additional features (e.g., intermediate data caching)
- ▶ Learn from other standardisation processes (e.g., ISO)
- ▶ Regular revision cycles every 5 years or alternatively establish constantly available system (“complaint box”)
- ▶ OECD to initiate and co-ordinate revision cycles

Feedback from the workshops with regard to the project results:

- ▶ Further in-depth evaluation of the received comments and proposals is needed to decide on the need for further TG revision processes
- ▶ Desire of (scientific) experts to be more involved in the OECD TG update/development process
- ▶ Define two pathways for OECD TG updates (1) for “general revisions” and (2) for minor adaptations”
- ▶ National-Co-ordinators are the contact point for experts intending to contribute to TG update process(es)
- ▶ Harmonization within the OECD TGs and also with other international standards is paramount to improve efficient and high-quality data generation (e.g., by means of OECD Guidance Documents or within the TGs themselves)

Following the six workshops, two further presentations were given to the National Co-ordinators involved in the OECD Test Guidelines Programme (WNT) to summarize the outcome of the project, and to lead an open discussion on the use of the results obtained for the future TG revision process. There was agreement that there is a high demand for updating certain TGs and efforts should be undertaken to establish a regular process to check update needs and to support the required revision process.

Zusammenfassung

Die OECD-Prüfrichtlinien (TG) stellen ein wichtiges Instrument zur Bewertung der Auswirkungen von Chemikalien auf die Umwelt dar. Die Aktualisierung der TGs wird von der Arbeitsgruppe der nationalen Koordinatoren für das OECD-Prüfrichtlinienprogramm (WNT) und dem OECD-Sekretariat organisiert, wobei sich das Verfahren auf Vorschläge der OECD-Mitgliedstaaten stützt und keine regelmäßige Aktualisierungsprüfung beinhaltet. Um sicherzustellen, dass die OECD TGs den Stand von Wissenschaft und Technik widerspiegeln und den regulatorischen Anforderungen der Mitgliedsländer entsprechen, sollen die OECD TGs kontinuierlich erweitert und aktualisiert werden. Da eine regelmäßige Überprüfung der OECD TGs nicht vorgeschrieben ist, liegt es in der Verantwortung der OECD-Mitgliedsländer, die für die Aktualisierung der TG erforderlichen Projekte zu ermitteln, vorzuschlagen und durchzuführen. Es ist daher nicht ungewöhnlich, dass nur diejenigen TG aktualisiert werden, an denen die Mitgliedsländer ein besonderes Interesse haben und für die genügend Ressourcen zur Verfügung stehen, um einen Überarbeitungsprozess einzuleiten. Es liegt auf der Hand, dass dies zu einer Vernachlässigung anderer TG führen kann, für die jedoch ebenso eine Überarbeitung notwendig sein könnte.

Das Umweltbundesamt (UBA) hat das Fraunhofer IME, das Fraunhofer ITEM und Ramboll beauftragt, einen möglichen Aktualisierungsbedarf für TGs zu identifizieren und zu priorisieren, die sich auf die Auswirkungen von Chemikalien auf biotische Systeme, auf das Verhalten und den Verbleib von Chemikalien in der Umwelt oder auf ihre physikalisch-chemischen Eigenschaften bei der Umweltrisikobewertung beziehen. Etwaiger Aktualisierungsbedarf sollte auf der Grundlage von Rückmeldungen aus der internationalen Gemeinschaft erhoben werden.

In einer groß angelegten Online-Umfrage von Mai bis Juli 2021 sammelte das Projektteam 257 ausgefüllte Fragebögen, die rund 500 individuelle Aktualisierungsvorschläge enthielten. Darüber hinaus wurden 59 Beiträge per E-Mail an das Projektteam geschickt und für das Projekt weiter berücksichtigt. Ursprünglich war vorgesehen, dass das Projektteam die Aktualisierungsvorschläge anhand entwickelter Kriterien wie Anzahl der eingegangenen Vorschläge, Jahr der letzten Überarbeitung, damit verbundener Zeitaufwand usw. priorisiert. Aufgrund der großen Anzahl von Rückmeldungen und der sehr hohen Teilnahmequote wurde jedoch eine zweite Umfrage - die so genannte Vorab-Evaluierungsumfrage - durchgeführt, in der die internationale wissenschaftliche Gemeinschaft mit den zuvor eingegangenen Aktualisierungsvorschlägen konfrontiert wurde. Indem das Projektteam den internationalen Expertinnen*Experten die Möglichkeit gab, alle vorgeschlagenen Aktualisierungsvorschläge für eine bestimmte TG einzusehen, konnte es anschließend um eine allgemeine Einschätzung der Relevanz einer Aktualisierung für die TG bitten. Beide Umfragen standen allen interessierten Parteien offen, einschließlich Industrie, Wissenschaft, Vertragslaboren, NROs und Behörden.

Die Relevanz der vorgeschlagenen Überarbeitungen der verschiedenen TGs wurde vom Projektteam auf drei verschiedene Arten bewertet. Erstens wurden die TGs nach der Anzahl der in der Online-Umfrage eingegangenen Aktualisierungsvorschläge sortiert. Zweitens wurden diejenigen TGs ermittelt, die von der Mehrheit der Teilnehmenden in der Vorbewertung hinsichtlich der Relevanz einer Überarbeitung als hoch eingestuft wurden (ausgedrückt in %). Schließlich wurde die Bewertung der einzelnen TG (geringe/mittlere/hohe Revisionsrelevanz) durch Gewichtung der Ergebnisse weiter verfeinert. Basierend auf den Ergebnissen der verschiedenen Ranking-Ansätze wurde die Revisionsrelevanz der einzelnen Leitlinien priorisiert.

In einem weiteren Ansatz wurde die für jede Leitlinie ermittelte Relevanz der TG-Revision (gewichtet in %) mit dem Jahr der Veröffentlichung der TG verglichen. Die Ergebnisse zeigen,

dass keine Korrelation zwischen den beiden Faktoren gefunden wurde, auch wenn nur die am höchsten bewerteten TGs berücksichtigt wurden.

Die Teilnehmenden der Vorabbewertung der Vorschläge wurden nach ihrem Interesse an der Mitarbeit in einer potenziellen OECD-Expertengruppe befragt, die sich mit der Überarbeitung einer bestimmten TG befasst. Im Allgemeinen wurde ein hohes Interesse an der Mitarbeit in einer OECD-Expertengruppe festgestellt. Die Ergebnisse wurden mit der ermittelten Relevanz einer Überarbeitung für die verschiedenen TGs verglichen. Interessanterweise stießen die TGs, für die die höchste Relevanz einer Überarbeitung ermittelt wurde, nicht unbedingt auf das größte Interesse bezüglich der Teilnahme an einer entsprechenden OECD-Expertengruppe. Dies könnte zumindest teilweise durch den begrenzteren Kreis von Experten und Expertinnen bei sehr spezifischen TGs (z.B. Vogeltoxizität) im Vergleich zu allgemein genutzten TGs (z.B. Algen- oder akute Fischtoxizität) erklärt werden.

Mit dem Projekt sollte überprüft werden, für welche OECD-TGs ein Überarbeitungsbedarf besteht, weil sie nicht dem aktuellen Stand von Wissenschaft und Technik entsprechen. Es wurden alle TGs überprüft, die für die Bewertung der Auswirkungen von Chemikalien auf die Umwelt relevant sind. Die folgenden Schlussfolgerungen können gezogen werden:

- ▶ Eine hohe Anzahl von Vorschlägen zur weiteren Verbesserung der OECD TGs ist eingegangen (max. 55 pro TG)
- ▶ Eine hohe Anzahl von Befragten pro TG (5-35) trug zur Bewertung der allgemeinen Relevanz einer weiteren Überarbeitung der einzelnen TGs bei
- ▶ Vorab-Bewertung der Vorschläge: Vielfältige Kommentare, die Zustimmung/Ablehnung zu den Aktualisierungsvorschlägen ausdrücken, liegen vor und bieten eine fundierte Grundlage für die weitere Diskussion und einen zukünftigen TG-Aktualisierungsprozess
- ▶ TGs mit hoher Relevanz einer zukünftigen Überarbeitung wurden identifiziert.
 - TOP 3 “Verbleib und Verteilung in der Umwelt”
 1. Test No. 301: Ready Biodegradability
 2. Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure
 3. Test No. 310: ready Biodegradability – CO₂ in sealed vessels (Headspace test)
 - TOP 3 “Terrestrisch”
 1. Test No. 208: Terrestrial Plant Test: Seedling Emergence and Seedling Growth Test
 2. Test No. 227: Terrestrial Plant Test: Vegetative Vigour Test
 3. Test No. 222: Earthworm Reproduction Test (*Eisenia fetida*/*Eisenia andrei*)
 - TOP 3 “Aquatisch”
 1. Test No. 231: Amphibian Metamorphosis Assay
 2. Test No. 201: Freshwater Alga and Cyanobacteria, Growth Inhibition Test
 3. Test No. 234: Fish Sexual Development Test
- ▶ Es hat sich gezeigt, dass die Bedeutung der Überarbeitung unabhängig vom Erscheinungsjahr einer TG ist
- ▶ Es besteht ein großes Interesse an der Teilnahme an OECD-Experten*Expertinnen Gruppen, um den weiteren TG-Überarbeitungsprozess zu unterstützen

Nach Abschluss der Umfragen wurden die Ergebnisse mit allen interessierten Parteien, einschließlich Industrie, Wissenschaft, Vertragslabors, NGOs und Behörden, diskutiert. Es wurden drei virtuelle Workshops durchgeführt, die jeweils einen der drei verschiedenen Blöcke ("Verbleib und Verteilung in der Umwelt", "Terrestrisch" und "Aquatisch") abdeckten. Jeder Workshop wurde an zwei verschiedenen Tagen zu unterschiedlichen Zeiten angeboten. Wichtige Aspekte, die sich aus der Erhebung und der Vorbewertung ergaben, wurden erörtert, und es war genügend Zeit vorgesehen, um über die Erfahrungen bei der Durchführung der Erhebung zu berichten.

Das Feedback aus den Workshops im Hinblick auf ein zukünftiges Verfahren für den Aktualisierungsbedarf von OECD TGs war:

- ▶ Erweiterung zur besseren Einbindung von Experten aus dem außereuropäischen Ausland sowie von Experten aus der Wissenschaft und von NGOs
- ▶ Verlängerte Kommentierungsfristen, um eine sorgfältige Überarbeitung der großen Anzahl von TGs zu ermöglichen
- ▶ Verbesserte Website-Einstiegsseite mit zusätzlichen Funktionen (z.B. Möglichkeit zur (Zwischen-)speicherung)
- ▶ Lernen von anderen Normungsprozessen (z.B. ISO)
- ▶ Regelmäßige Revisionszyklen alle 5 Jahre oder alternativ Einrichtung eines ständig verfügbaren Revisions-Systems ("Beschwerdebox")
- ▶ OECD sollte Revisionszyklen initiieren und koordinieren

Rückmeldungen aus den Workshops im Hinblick auf die Projektergebnisse:

- ▶ Weitere eingehende Bewertung der eingegangenen Kommentare und Vorschläge ist erforderlich, um über die notwendigen TG-Überarbeitungsprozesse zu entscheiden
- ▶ Wunsch der (wissenschaftlichen) Experten*Expertinnen, stärker in den Prozess der Aktualisierung/Entwicklung der OECD TG einbezogen zu werden
- ▶ Definition von zwei Wegen für OECD TG-Aktualisierungen (1) für "allgemeine Überarbeitungen" und (2) für kleinere Anpassungen (verkürzt)
- ▶ Nationale Koordinatoren*Koordinatorinnen sind die Kontaktstelle für Experten*Expertinnen, die einen Beitrag zum TG-Aktualisierungsprozess leisten wollen
- ▶ Harmonisierung innerhalb der OECD TGs und auch mit anderen internationalen Standards ist von größter Bedeutung, um eine effiziente und qualitativ hochwertige Datengenerierung zu gewährleisten (z.B. mittels OECD Guidance Documents oder innerhalb der TGs selbst)

Im Anschluss an die sechs Workshops wurden zwei weitere Präsentationen für die am OECD-Prüfungsrichtlinienprogramm (WNT) beteiligten nationalen Koordinatoren*Koordinatorinnen gehalten, um die Ergebnisse des Projekts zusammenzufassen und eine offene Diskussion über die Verwendung der erzielten Ergebnisse für den künftigen TG-Überarbeitungsprozess zu führen. Es herrschte Einigkeit darüber, dass ein hoher Bedarf an einer Aktualisierung bestimmter TG besteht und dass Anstrengungen unternommen werden sollten, um ein regelmäßiges Verfahren zur Überprüfung des Aktualisierungsbedarfs und zur Unterstützung des erforderlichen Überarbeitungsprozesses zu schaffen.

1 Background and objectives

Key elements

- ▶ OECD Test Guidelines display an important tool to evaluate the impacts of chemicals on the environment.
- ▶ OECD Test Guidelines are the basis of the OECD-wide system of mutual acceptance of data.
- ▶ Updates rely mainly on proposals from OECD member countries.
- ▶ There is no regular update check.
- ▶ The German Environment Agency contracted Fraunhofer IME, Fraunhofer ITEM and Ramboll to identify and prioritise any potential update needs and to come up with a list for potential actions. It is emphasised that the project shall rely on feedback from the international community.

OECD Test Guidelines (TGs) display an important tool to evaluate the impacts of chemicals on the environment. The update of the TG is organised by the Working Party of National Co-ordinators of the OECD Test Guidelines Programme (WNT) and the OECD Secretariat. The procedure relies on proposals from OECD member countries and does not contain a regular update check.

The German Environment Agency (UBA) contracted Fraunhofer IME, Fraunhofer ITEM and Ramboll to identify and prioritise any potential update needs. It was emphasised that the project shall rely on feedback from the international community.

The research project was intended to review which OECD TGs do not correspond to the current state of science and technology. All TGs relevant for the assessment of effects on the environment were to be reviewed. This included all TGs from areas 2 (effects on biotic systems) and 3 (behaviour and fate to the environment) and partly TGs from area 1 (physicochemical properties), which are relevant for the evaluation of effects of chemicals on the environment (e.g., octanol-water coefficient, hydrolysis, water solubility).

The main objective of the project was to provide a list of potential update needs related to OECD TGs based on input from the international scientific and regulatory communities. In order to facilitate any follow up activities, the project team was asked to evaluate the proposed update needs for prioritisation. It is then up to the OECD member countries to take any further actions on updating the specific TGs based on the knowledge gained through this project.

To incorporate the opinion of international scientific experts, a survey was conducted. The questionnaire was sent to international experts and stakeholders to gain an overview of the requests for updates and changes of OECD TGs. Participants in the survey had the opportunity to comment and propose changes on single or multiple OECD TGs relevant to environmental assessment. In parallel, prioritisation criteria were elaborated, which should help to prioritise the provided update suggestions later on. The prioritisation approach has been discussed and further elaborated with selected experts in a workshop in September 2021, what led to a second survey (in the following called pre-evaluation) to involve the international community in the assessment of the responses.

Based on the results of the first survey and the responses to the pre-evaluation, several international workshops were conducted to discuss the survey results and proposed updates related to environmental aspects in OECD TGs.

2 Project approach

In a nutshell

What is the aim of the project?

To check whether existing OECD Test Guidelines (from area 2, 3 and partly 1) correspond to the current state of science and technology and to identify any related update needs

The project is not aimed at:

- ▶ Actual update of any Test Guideline
- ▶ Binding update suggestions
- ▶ Any obligations for anybody

What should be the main output of the project?

- ▶ A list of identified update suggestions
- ▶ Accompanied by a first assessment of importance, workload and demand

How were the update needs identified?

Open online survey to receive input from the international community on update needs

As explained in the background chapter, the project should not concentrate on a specific TG to be updated but should systematically identify all update needs in relation to the TGs relevant for the environment risk assessment (ERA).

In order to identify the update needs an extensive online survey has been conducted, for which access was made available via the project website¹. The respondents were guided along the structure of the different TGs in order to enable the project team to precisely identify the paragraph for which the update is relevant. All respondents were asked to also provide a rationale for each proposed revision to make these easier to understand.

Before the survey started the project team elaborated several criteria which were foreseen for prioritising the incoming suggestions. As will be explained in the following chapters, this approach however, was not sufficient as the proposed update suggestions were often technically very detailed and brought up only by a single respondent. It was therefore difficult to judge if the update suggestions contained a valid point. In some cases, the correctness and meaningfulness of the proposed update suggestion could only have been ensured by verification in the laboratory or by expert judgement that could not be provided by the project team alone.

Therefore, an online pre-evaluation was conducted on all received revision proposals which was open to all experts/stakeholders that had already been involved in the first survey. The respondents were asked to rate the necessity and urgency of each revision proposal. Further details are available in the following chapters.

Following the pre-evaluation step several expert workshops were conducted in order to discuss results of the study, potential next steps.

¹ <https://rev-otg.com>; remains available until 2025

3 Data collection

3.1 Online survey (May 2021 – July 2021)

Key elements

- ▶ Link to online survey was publicly available
- ▶ Advertisement during SETAC and via email invitations
- ▶ Questionnaire guided respondents along the structure of the TGs
- ▶ Additional responses possible in any format
- ▶ Survey was open May – July 2021.

3.1.1 Information needs

Information needs could be allocated to three categories, namely: (i) precise information on what needs to be updated, (ii) information enabling prioritisation and (iii) some additional cross check information.

For aspect (i) the project team identified the following necessary aspects:

- ▶ Precise location (chapter, subchapter) of the update
- ▶ Rational – why is the update needed?

For this reason, the project team transferred the main structure of each relevant OECD Test Guideline into the questionnaire so that the respondent could easily click along the chapters and subchapters. As soon as the respondents had provided the information on the chapter or subchapter, he/she wanted to comment on, open text fields were available to explain the update suggestion and the rational.

For aspect (ii) the project team identified some of the prioritisation criteria originally considered, to be asked of the respondent as well. This was in particular:

- ▶ Relevance of the suggested update for regulatory purpose
- ▶ General relevance of the reason for revision
- ▶ Type of revision required

In the questionnaire, answer options in this regard were predefined so that the respondent just needed to click on the preferred answer.

And finally, for aspect (iii) the cross-check information the following information needs had been identified:

- ▶ information on the background of the respondent (expertise level and way of working with Test Guidelines)

Also here, some predefined answer options were available to support the respondents.

3.1.2 Conducting the online survey

The survey has been prepared with Ramboll's SurveyXact tool and could be accessed via any common internet browser on a computer or mobile equipment. The questionnaire was programmed as an anonymous survey, meaning that no other data was collected about the respondent other than the questions indicated. This limited some technical functions but ensured full data privacy for any participating expert. In addition, a pdf version was made available via the website and experts were encouraged to use any means of communication to report known TG update needs.

As far as possible automatic, logical jumps were programmed, meaning that irrelevant questions for the respondent were skipped automatically. For one or more update suggestions relevant for a specific TG one questionnaire could be used, for answers relevant for another TG the respondent was directed to another questionnaire. This was done to facilitate the automated assessment of the provided answers that would not have been possible with mixed questionnaires.

The link to the survey has been distributed via various mailing lists (covering laboratory, regulatory and scientific experts), supported by UBA and OECD.

A project website has been established to advertise the project and the survey and is available on: <https://rev-otg.com>. Separate pages for background and aim of the project, procedures, survey and workshop have been established. Furthermore, a contact page and the obligatory imprint was provided.

Some overall numbers

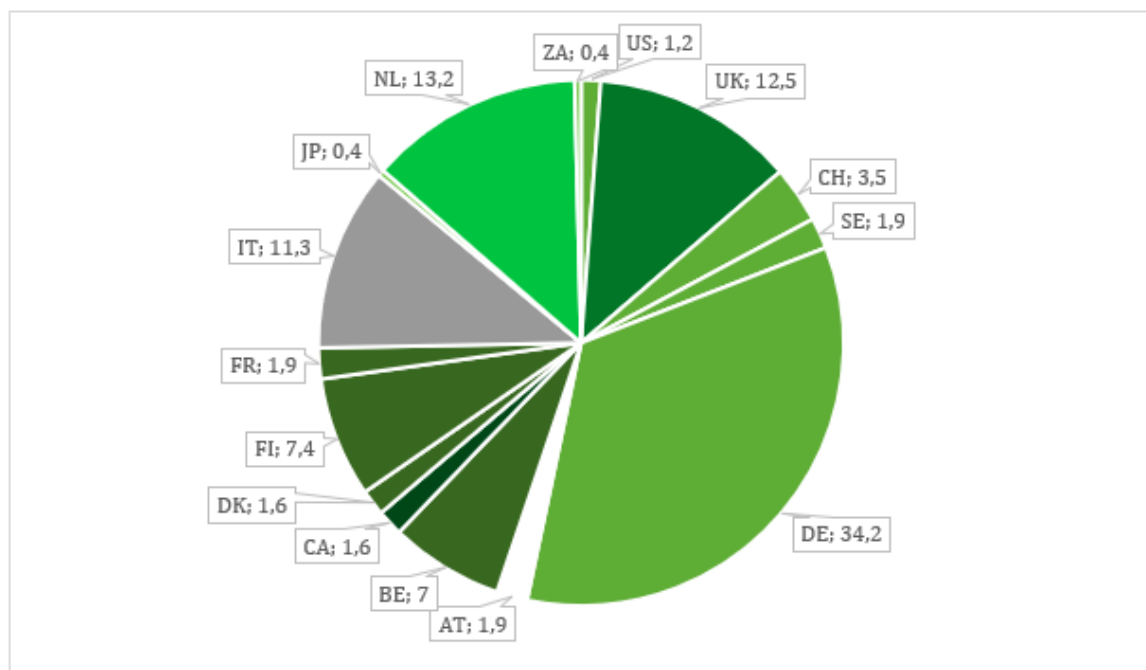
- ▶ 257 completed questionnaires
- ▶ ~ 500 individual suggestions via questionnaires
- ▶ 59 additional comments via email

Overall, 257 completed questionnaires could be obtained, 59 additional comments were received via email summing up in a total number of more than 500 individual update suggestions provided by the respondents. The following figures provides information on the geographical distribution as well as the distribution among different stakeholder groups.

The majority of responses were received from by German experts (34.2%), British experts (12.5%), the Dutch experts (13.2%) and Italian experts (11.3%). Overall, experts from 15 countries contributed to the survey. Considering the responder groups, major input came from authorities (47%), industry (22% by industry, 3% by industry associations) and contract research organisations (12%),

Figure 1: Geographical distribution of respondents

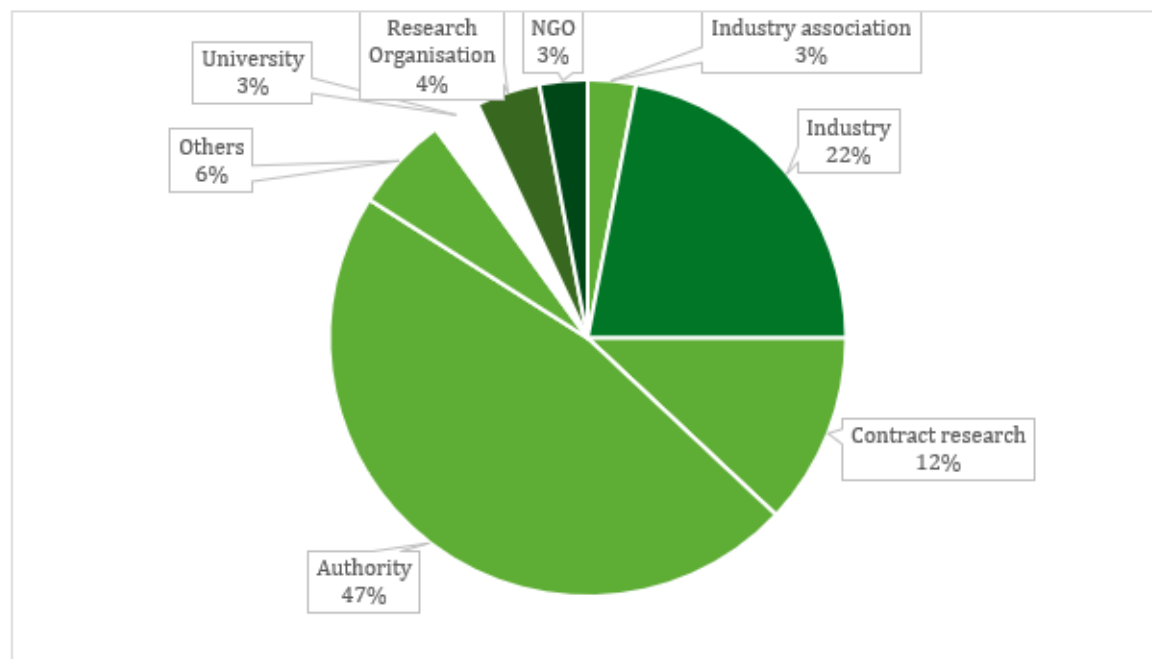
Countries (DE= Germany, SE= Sweden, CH= Switzerland, UK= United Kingdom, US=United States of America, ZA=South Africa, NL=Netherlands, JP= Japan, IT= Italy, FR= France, FI= Finland, DK=Denmark, CA=Canada, BE= Belgium, AT= Austria) and percentage



Source: own illustration, Ramboll

Figure 2: Distribution for responder groups

Kind of stakeholder and percentage



Source: own illustration, Ramboll

3.2 Pre-evaluation (Nov 2021 – Jan 2022)

Originally, it was foreseen to evaluate the collected suggestions using different assessment criteria. These were discussed, selected and defined before the survey started. The criteria were:

- ▶ Criterion 1: Time since last update
- ▶ Criterion 2: Number of feedbacks received to a specific update
- ▶ Criterion 3: Relevance of the TG for regulatory purpose
- ▶ Criterion 4: General relevance of the reason for revision
- ▶ Criterion 5: Type of revision required
- ▶ Criterion 6: Necessary revision depth
- ▶ Criterion 7: Effort of revision

After having elaborated a clear understanding of each criterion itself, it was defined how the criterion should be used in the prioritisation process. In principle, a ranking exercise following a scoring system was foreseen to prepare a list of update needs to be ranked according to their score assessed according to the prioritisation criteria.

However, when the update suggestions became available it turned out that for a majority of these it was not possible to assess the defined criteria without consultation of additional experts for the specific TG because of the technical level of detail provided.

Against this background a team of experts, consisting of participants from German Environment Agency (UBA), the German Federal Ministry of the Environment, Nature Conservation, Nuclear Safety and Consumer Protection, the German Federal Institute of Risk Assessment (BfR) as well as from Fraunhofer Institute and OECD discussed and tested in a workshop in September 2021 the possibility to pre-evaluate the received update suggestions. After a general introduction to the project and an overview on received update suggestions, all experts received a document, which contained all received update suggestions sorted by the individual TGs. All experts were asked to have a look on the TGs for which they have expertise, so that they were aware of the full range of suggested updates. They were then asked to rate the relevance of the proposed guideline revision based on the categories low, medium or high. The approach was fine-tuned afterwards in order to enable also individual remarks for specific update suggestion. Finally, the approach was programmed via an online questionnaire to make use of the knowledge and experience of the broader scientific community to get feedback on the usefulness and necessity of the suggested updates.

In order to direct respondents for the pre-evaluation directly to the TGs they have expertise in the project team allocated all guidelines to three different blocks (aquatic, terrestrial and fate & distribution).

For each block a different questionnaire was prepared. In these questionnaires all received update suggestions were made publicly available. Subsequently, experts were asked whether they agree or disagreed with the proposed updates, and additional justification could be provided.

The experts were asked to evaluate the general relevance of the overall revision of each TG based on the individual revision proposals. Three response options were given:

- ▶ Low - meaning update addresses aspects that belong to the following categories: 'Spelling errors with no relevance' or 'Incorrect information without influence on the test result(s)'
- ▶ Medium – meaning the update addresses aspects that belong to the following categories: 'Explanatory information that is helpful but not necessarily needed; test result(s) is(are) not influenced'
- ▶ High - meaning the update addresses aspects that belong to the following categories: 'Spelling errors with relevance for the test and the test result(s)' or 'Missing information that is however necessary to conduct the test' or 'Incorrect or missing information with relevance for the test result(s)' or 'Update suggested reflects more recent state of the art techniques & approaches'

In addition, all experts were asked whether they would be interested in participating in an OECD Expert Group if an OECD update process were to be initiated in the future. The results of the pre-evaluation are shown in Appendices A-C.

4 Key results – Suggestions and Pre-evaluation

As mentioned above, the survey led to a high number of completed questionnaires (257), and additional comments via email (59). These submissions often included several individual suggestions for OECD TG improvements, so that the number of update suggestions was far higher. Redundancies in the list of suggestions were rather rare.

In Figures 3-5, the number of questionnaires and the number of update suggestions is illustrated for each OECD TG, which were grouped to the three different blocks (“fate”, “terrestrial” and “aquatic”). Table 1 provides for each block a summary of the number of submitted update suggestions. If no update suggestions were received for a specific TG, the document was not included in the overview.

For the “Fate” block more than 10 update suggestions were submitted for OECD TGs 301, 307, 308, 309, and OECD TG 106. This group also included the TG with the highest number of update suggestions (> 40, OECD TG 309). Between 20-40 update suggestions were submitted for three TGs. In the “Terrestrial” block more than 10 update suggestions were received for OECD TGs 206, 208, 213, 214, 222, 226, 227 and 232, with the number of proposals ranging from 20 to 40 for six TGs and not exceeding 40 for any TG. For the “Aquatic” block more than 10 update suggestions were received for OECD TGs 201, 202, 203, 210, 211, 215, 221, 229, 231, 234, and 236, but only for two TGs between 20 and 30 update suggestions (none > 40) were obtained.

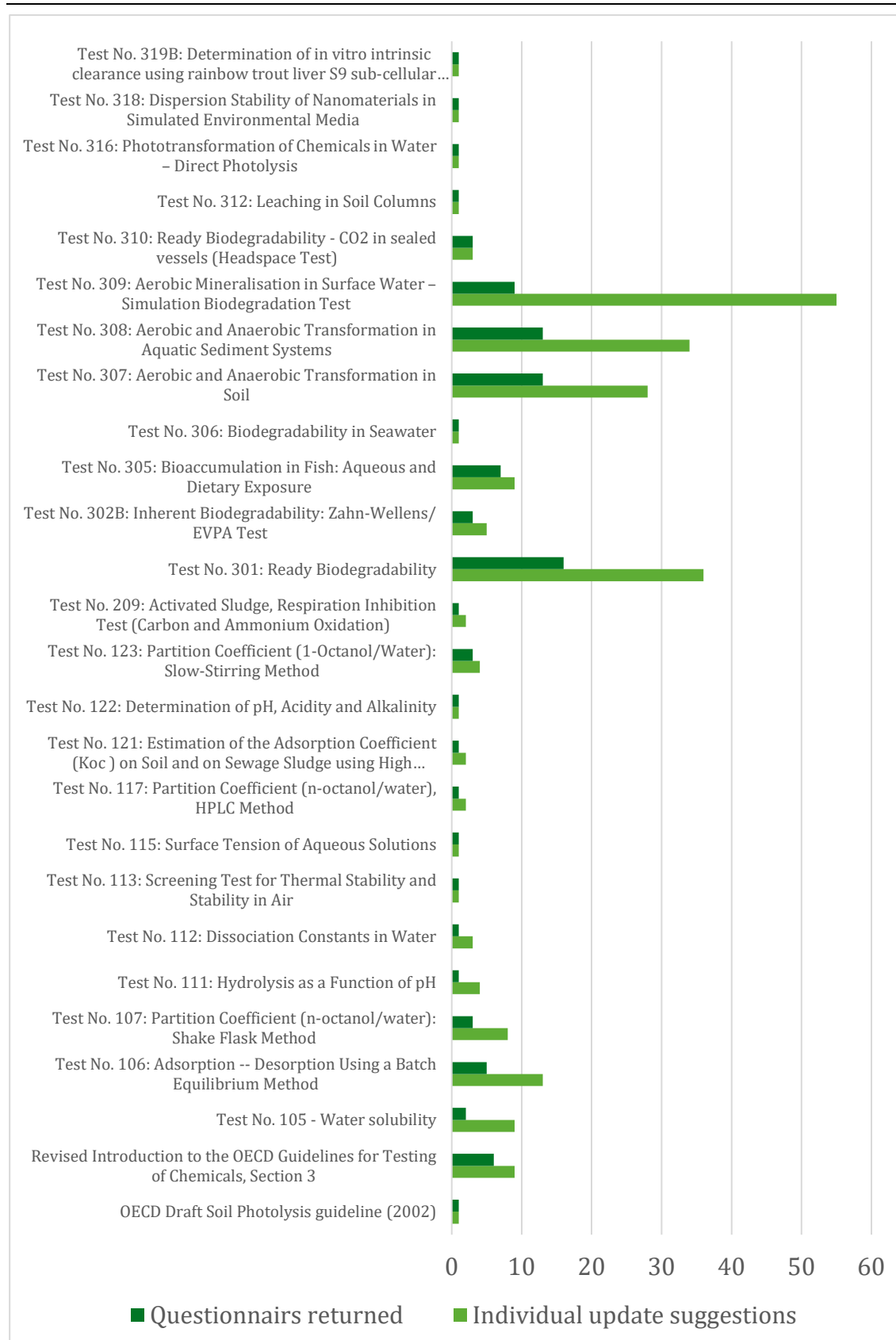
Looking closer at the TGs that have received a high number of comments, the overall impression is that TGs with higher regulatory relevance received more update suggestions.

Table 1: Number of update suggestions

	Total No of TGs with update suggestions	No of TGs with update suggestions of < 10	No of TGs with update suggestions of 10-20	No of TGs with update suggestions of 20-30	No of TGs with update suggestions of 30-40	No of TGs with update suggestions of > 40
Fate	26	21	1	1	2	1
Terrestrial	19	11	2	4	2	-
Aquatic	25	14	9	2	-	-

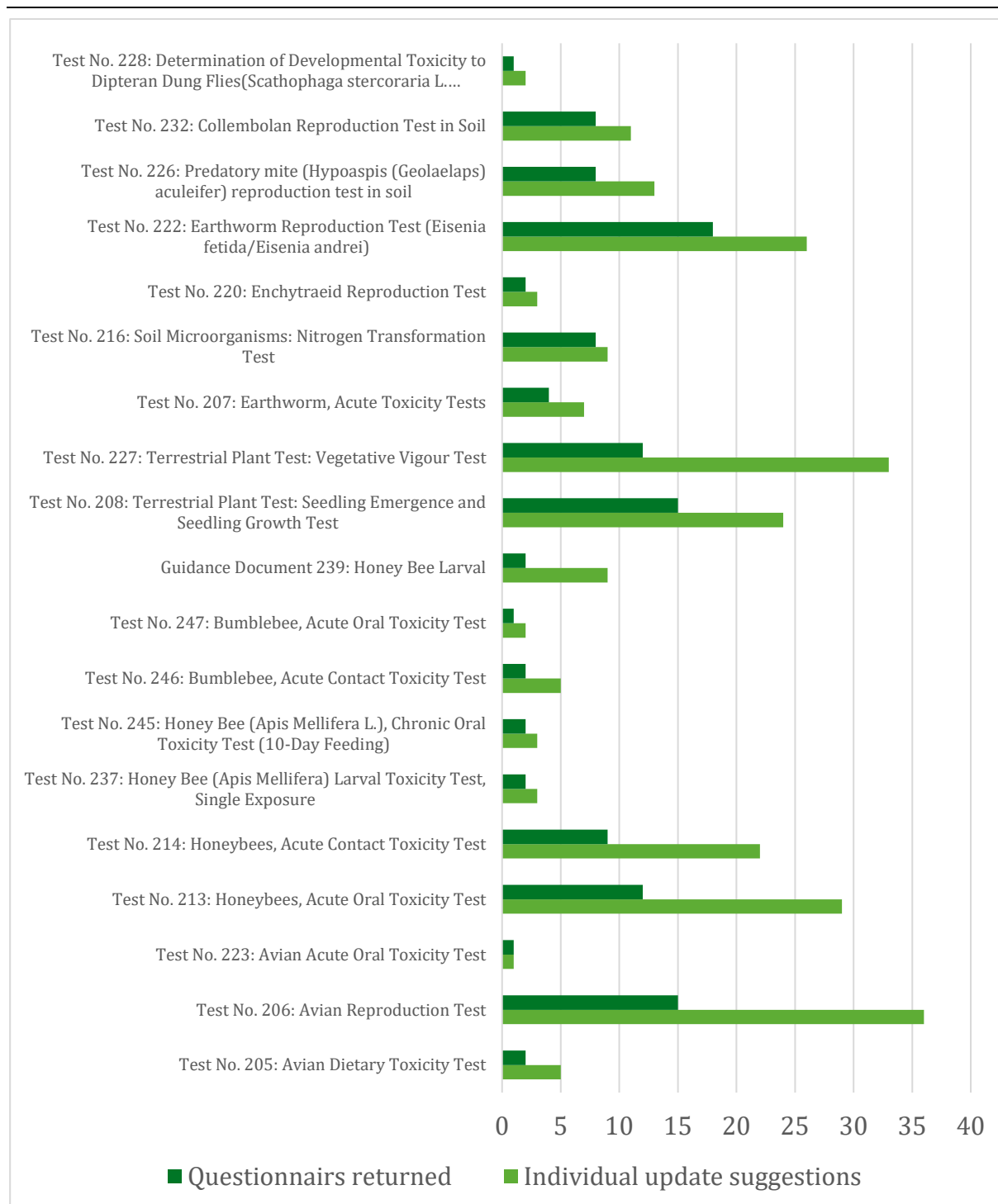
The number of suggestions per TG may indicate the relevance of a revision, and thus can help to optimise the prioritisation process. However, they do not allow a final conclusion on the requirement for a further revision. For this reason, in Chapters 4.1 and 4.2 a grouping of the different types of update suggestions and the results of the pre-evaluation approach regarding the relevance of revision are presented, respectively.

Figure 3: Survey: Number of returned questionnaires and individual update suggestions “Fate”



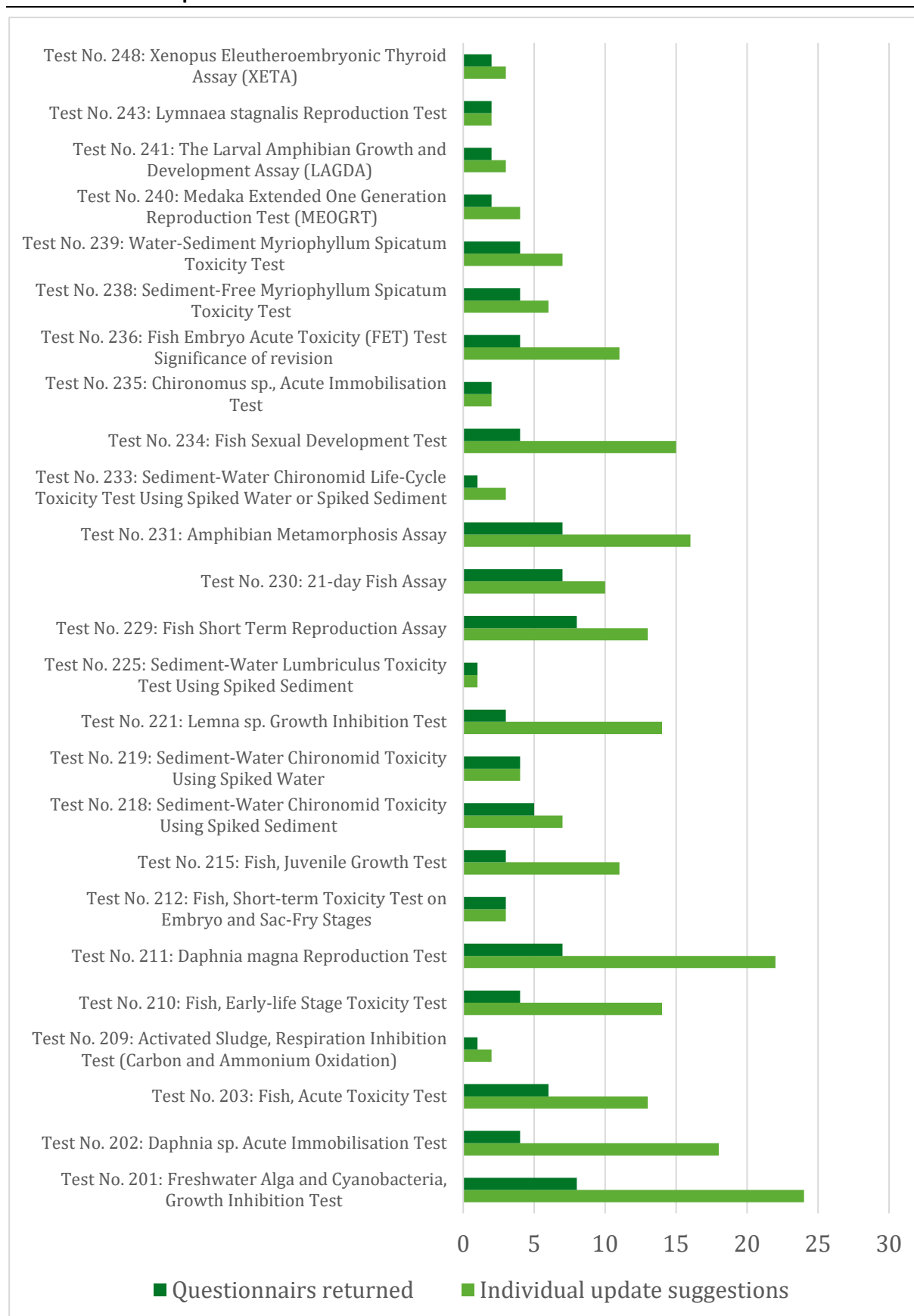
Source: own illustration, Ramboll

**Figure 4: Survey: Number of returned questionnaires and individual update suggestions
“Terrestrial”**



Source: own illustration, Ramboll

Figure 5: Survey: Number of returned questionnaires and individual update suggestions “Aquatic”



Source: own illustration, Ramboll

4.1 Type of update suggestions

All update suggestions received during the survey are presented in the Appendices A-C to this report. The suggestions may have a different relevance for test conduction or evaluation and might even have an influence on the test results obtained when performing the TG. Depending on the complexity of the issues addressed in the survey, different efforts may be required within a revision process. However, even proposed changes which are editorial only at a first glance, can have vast consequences with regard to the different aspects mentioned (e.g., test conduction, results, reporting, etc.).

Overall, the update suggestions can be roughly grouped into the following types:

- ▶ General
- ▶ Editorial
- ▶ Scientific specific
- ▶ Relevance for test conduction or evaluation
- ▶ Relevance for regulation

In the following, examples for these types of update suggestions are provided to give an impression of the variety of suggestions. However, it should not be overlooked that some suggestions could be grouped in other ways.

4.1.1 Type of suggestions – General

Often the update suggestions received related not only to one specific TG but to several TGs and thus concerned harmonization issues between different OECD TGs. The issues mentioned are for example, comparability of test results of a certain group of TGs, analytical verifications, statistical methods, guidance on applicability of difficult to test substances or specific groups of substances, etc. In addition, harmonization issues were addressed with respect to other guideline frameworks (e.g., United States (US) OPPTS (Office of Prevention, Pesticides & Toxic Substances) and OCSPP (Office of Chemical Safety and Pollution Prevention, since 2010) guidelines, International Organization for Standardization (ISO), etc.).

Examples for type of suggestions - General

Note: All received update suggestions are provided in the appendices. In the following textbox, only examples for general suggestions are presented in the original wording including submitted rationales. No rationale is provided why the suggestions are grouped as 'general' or whether the suggestion is justified.

Revised Introduction to the OECD Guidelines for Testing of Chemicals, Section 3, para 13

It was suggested that "...with the OECD 301/302/306 & 310 TGs in their current form the following sentence does not generate comparative data that are used to derive the same conclusions." Therefore, it was recommended to perform further work (please see the rationale), "Ready biodegradability tests must be designed so that positive results are unequivocal".

TG 203 (Fish, Acute Toxicity Test), para 24

"It would be useful to clarify how to consider the information from the **analytical** measurements for the interpretation of the test. Note that this comment is relevant for all aquatic tests. However,

the different TGs give different level of detail on that issue; basically, a good harmonization would be needed...”

TG 213 (Honeybees, Acute Oral Toxicity Test), para 19

“It should be mentioned, that for microbials the test duration of several hours (max. 96 h) is too short, to assess pathogenic and infective properties of the test substance. Similar to **OPPTS 885.4380** the test duration could be prolonged up to 30 days provided that the control mortality does not exceed 20%.”

TG 205 (Avian Dietary Toxicity Test), Table 1

“Revise age of birds, test species, and test conditions to comply with **US guidelines**”

TG 206 (Avian Reproduction Test)

“Compared to other test systems like for mammals, a considerably reduced set of parameters is included. The development of a **completely new guideline**, in the best case with a 2-generation test design, should be considered.”

TG 301 (Ready Biodegradability), para 9

“As a ready biodegradability test (RBT) provides key data for many substance registrations, further development and guidance are needed to improve ready biodegradability tests of **difficult to test substances**. Recent academic publications have highlighted advances in passive dosing of hydrophobic substances to these test systems with positive results.”

TG 230 (21-day Fish Assay)

“The methodology as described in the current guideline is not suitable to test for effects of virtually **insoluble particulate materials** (e.g., nano-forms of substances, polymers). Adaption of the guidelines methodology with latest scientific findings is indicated so that valid tests with particulate materials can be performed.”

TG 206 (Avian Reproduction Test)

“Define workflow and principle of the **statistical tests**, which should include outlier identification, and how to decide suitability of trend tests.”

Requirement for Guidance Documents (GDs): Several “general” update suggestions were relevant for more than one or all TGs. Therefore, the question arose whether these update suggestions should be dealt with in higher-level guidance documents rather than within the individual TGs. The “Revised introduction to section 3 of the OECD guidelines for Testing of Chemicals” can be seen as such guidance. However, sometimes it needs to be clarified whether the proposed task can be addressed under the OECD TGP or under the interpretation of independent chemical regulations of individual OECD member countries, such as the EU REACH Regulation.

Harmonization between OECD TGs: Similar proposals for harmonization between OECD TGs were put forward regarding the preparation of test solutions of difficult-to-test substances as well as insoluble particulate materials. Also, the applicability of individual TGs was addressed in this context. An OECD GD² is already available for difficult to test substances in aquatic toxicity

² [Guidance Document on Aquatic Toxicity Testing of Difficult Substances and Mixtures | en | OECD](#)

tests. However, related issues are also the case for fate or terrestrial test systems and the guidance should be extended, respectively. For the statistical analysis some guidance is available (OECD Series on Testing and Assessment no. 54 Current approaches in the statistical analysis of ecotoxicity data: a guidance to application³, 2006)] but possibly needs more harmonization between OECD TGs, and additionally an overarching GD is required.

Harmonization between different guideline frameworks: Several general suggestions were related to the harmonization of OECD TGs with different guideline frameworks such as US guidelines (OPPTS/OCSP), European guidelines (EN - European Standards), or Japanese Guidelines.

New TGs: Although out of the scope of this review process, another general point raised is the potential need of new TGs as the required endpoints are not covered by the existing TGs and may not be solved by adaptation.

4.1.2 Type of suggestions – Editorial

Editorial changes seemed to be the easiest type of update suggestions to be solved. The suggested revisions addressed any editorial changes that do not have any influence on test results. This category included misspelling, harmonization of units, invalid or outdated references, etc.

Examples for type of suggestion - Editorial

Note: All received update suggestions are provided in the appendices. In the following textbox, only examples for editorial changes suggestions are presented in the original wording including submitted rationales. No rationale is provided why the suggestions are grouped as 'editorial' or whether the suggestion is justified.

TG 106 (Adsorption -- Desorption Using a Batch Equilibrium Method)

"The Guideline uses "cm³" all through the document. System International d'unites (SI) recommend "ml".

TG 106 (Adsorption -- Desorption Using a Batch Equilibrium Method), para 87 - (Page 16, line 7)

"Freudlich adsorption coefficient" is misspelled and should read "Freundlich adsorption coefficient".

TG 222 (Earthworm Reproduction Test), para 39

"In the first sentence the words "and cocoon numbers" should be deleted."

TG 202 (Daphnia sp. Acute Immobilisation Test), para 4

Note: Guidance for testing substances with physical chemical properties that make them difficult to test is provided in a separate document (4). "Literature - reference (4) and corresponding text in paragraph 4 should be updated".

Editorial changes may be easy to implement, but on the other hand often seem to be not urgent. However, even editorial changes may have a strong impact on the reliability of the results, e.g., if deletion of parts have been suggested, or if an outdated reference is used requiring different settings. In specific cases editorial changes may be difficult to judge without specific knowledge

³ MONO 2006 18 - GD Statistics - Final Apr06.doc (oecd.org)

on the respective TG and the regulatory impact. For this reason, the decision whether an update suggestion is editorial or actually scientific specific or even of regulatory relevance was not always obvious, and thus might need to be further clarified.

4.1.3 Type of suggestions – Scientific specific

Update suggestions which are not only editorial, but which need in depth knowledge on the scientific state-of-the-art could be grouped as “scientific specific”. The update suggestions addressed any scientific change that is related to specific aspects, like a change in a method (e.g., analytics, statistics, new approaches).

Examples for Type of suggestion – Scientific specific

Note: All received update suggestions are provided in the appendices. In the following textbox, only examples for scientific specific suggestions are presented in the original wording including submitted rationales. No rationale is provided why the suggestions are grouped as ‘scientific specific’ or whether the suggestion is justified.

TG 106 (Adsorption -- Desorption Using a Batch Equilibrium Method),

para 95

Test report: “There are some new ideas on how to derive more “unit equivalent” Kf values (‘Kf’) by expressing the sorption isotherm with the equation $CS = 'KfCr_n$, where Cr (“reduced concentration”) is the equilibrium solution concentration normalised to the (subcooled) liquid aqueous solubility (Carmo et al., 2000). We are interested in this approach because it could give more “inter-compound comparable” data. Although the OECD guideline 106 never has recommended comparing traditional Kf values with different 1/n, this is frequently done within risk assessment, but is clearly incorrect since they have different units (like comparing m² with m³) and sometimes misleading. Have the ideas by Carmo et al., or similar approaches, been discussed by the Guideline work group?”

Para 86

Koc or better Doc: “It might be relevant to change to the notation log Doc (log organic carbon – water distribution ratio). This is used in modern textbooks on environmental organic chemistry (e.g. Schwarzenbach et al., 2003, page 321) in order to underline that it [is] a distribution ratio of (possibly) many sorbed and dissolved species. They use the same argumentation for why Dow should replace Kow (octanol – water). The current “Koc” values are clearly Doc in this sense, since they change with pH (indication the sorbing molecule species change composition).”

TG 239 (Water-Sediment Myriophyllum Spicatum Toxicity Test), para 89

“The paragraph addresses the controversy of the ErC50 vs the EyC50, which in itself is appreciated. However, it causes confusion with regard to certain publications, where the ErC50 has been compared to mesocosm results and found to be not always protective. Can the paragraph be revised in the light of these scientific findings?”

Update suggestions which are scientific specific need an in-depth discussion within an OECD Expert Group when these suggestions are checked for relevance. Thereby an agreement should be achieved on the scientific most adequate solution. However, it needs to be considered that this could have an impact on existing data, and thus an evaluation and description under which

conditions previous (“old”) data is still reliable might be necessary. Therefore, a transparent revision process based on a clear rationale is always required.

4.1.4 Type of suggestions – Relevance for test conduction or evaluation

Update suggestions with relevance for test conduction or evaluation were actually a sub-group of “scientific specific”. In addition, some points discussed under “general” may also be aspects for this category.

Examples for Type of suggestion – Relevance for test conduction or evaluation

Note: All received update suggestions are provided in the appendices. In the following textbox, only examples with relevance for test conduction or evaluation suggestions are presented in the original wording including submitted rationales. No rationale is provided why the suggestions are grouped as ‘relevance for test conduction or evaluation’ or whether the suggestion is justified.

GD 239 (Honey Bee Larval Toxicity Test following Repeated Exposure), para 11

Update suggestion: “Alternatively, the plates can be closed with the lid and transferred into an incubator at 34-35°C with a relative humidity within the range 50 - 80%. On day 22 the fully developed, adult honeybees are counted as emerged.”

TG 309 (Aerobic Mineralisation in Surface Water – Simulation Biodegradation Test)

“Introduce guidance on the best technique(s) to prepare the sterile control (309, 307 and if introduced for 308).”

TG 229 (Fish Short Term Reproduction Assay), Annex 2 / paras 18 and 23

“Aeration should be allowed since test start. Oxygen dropping below 60% renders the test invalid, however aeration is only permitted if oxygen drops below 60%.”

TG 206 (Avian Reproduction Test)

“Define workflow and principle of the **statistical tests**, which should include outlier identification, and how to decide suitability of trend tests.”

The same considerations as for the category „scientific specific“ are valid for this subgroup. In-depth knowledge is needed, and thus experts for the specific test systems should discuss if these update suggestions are justified. Thereby different TGs could be treated in one OECD Expert Group as some questions apply to several TGs even if not explicitly mentioned during the online survey.

4.1.5 Type of suggestions – Relevance for regulation

The update suggestions of this category are scientific specific but have at the same time a high relevance for regulatory decision making. The examples given below are probably influenced by current research activities and regulatory needs such as the need to evaluate endocrine disruptive properties.

Examples for Type of suggestion – Relevance for regulation

Note: All received update suggestions are provided in the appendices. In the following textbox, only examples for suggestions with relevance for regulation are presented in the original wording including submitted rationales. No rationale is provided why the suggestions are grouped to 'relevance for regulation' or whether the suggestion is justified.

TG 234 (Fish Sexual Development Test)

"It is suggested to add thyroid-specific endpoints to the guideline. Specifically: thyroid hormone measurements, thyroid histopathology, eye histopathology and assessment of swim bladder inflation."

TG 236 (Fish Embryo Acute Toxicity (FET) Test), paras 28, 34

"Since the use of the zebrafish embryo toxicity is already a basis to study additional endpoints, which are not related to acute toxicity, the extension of the observation, specifically observation on behaviour could be already included. One simple endpoint is for instance the touch evoked response. The test is easy to perform, by touching the embryos at stage 96 hpf with pipette tip or fine brush in the trunk region. Controls are showing a fast response, namely swimming away from the tactile stimulus."

TG 309 (Aerobic Mineralisation in Surface Water – Simulation Biodegradation Test)

"The test should include more validity criteria. The test substance, if not degraded, should be proven to have stayed in the water phase throughout the test (10% loss at the end of the test allowed max). If not, the test should be considered invalid. Also, the test should also give some range of acceptable degradation kinetics of the reference item in function of the test temperature. Oxygen concentration in water should not fall below 7 mg/L."

TG 206 (Avian Reproduction Test), para 11

"It should be clarified how the different endpoints can be compared and the dependency of some endpoints should be clearly described."

TG 227 (Terrestrial Plant Test: Vegetative Vigour Test), para 27

"It is suggested to introduce reproductive effects endpoints (e.g., flowering, seed production), as these sensitive effects on plants are not yet included in the risk assessment for non-target terrestrial plants. This will subsequently require updating of data requirements / guidance documents to enable a full plant risk assessment, covering the full lifecycle of plants. see Strandberg et al, 2019: Strandberg, B. et al, 2019. Pesticide effects on non-target terrestrial plants at individual, population and ecosystem level (PENTA), Danish Environmental Protection Agency, Pesticide Research 182, September 2019. available at <https://www2.mst.dk/Udgiv/publications/2019/09/978-87-7038-111-6.pdf>"

It is obvious that due to the relevance for regulation of this type of suggestions the need of the revision of the related TGs appears to be high and is likely leading to a high priority. However, also in this case the consequences of any revision including the continued validity of already existing data has to be evaluated. Guidance on how to deal with "older" data is necessary, especially when new validity criteria are introduced, i.e., instructions under which conditions

the "older" data, which were collected with the then obsolete TG without the new validity criteria, are still valid.

4.2 Pre-evaluation

As mentioned in chapter 4.1 it was not always obvious to allocate the received update suggestions to the different types of suggestion, defined within the project and it was even harder to oversee the relevance of a revision need. As these judgements have an impact on the priority setting for future revisions it was decided by the project team to have a pre-evaluation exercise asking all stakeholders to rate the general relevance of a certain TG revision based on the received update suggestions, and if wanted to provide an expert opinion on each update suggestion. This pre-evaluation exercise was again done as an online survey among all experts that were already contacted in the original survey. Overall, more than 400 (partially) completed questionnaires were returned in the pre-evaluation survey.

4.2.1 Pre-evaluation - rating of relevance of revision

In the second survey, participants (experts) were first asked to rate, based on the submitted update suggestions, the relevance of revision into the categories low, medium, and high for each individual TG. The results are presented and discussed in chapter 5 on relevance and prioritization.

4.2.2 Pre-evaluation - feedback on each single update suggestion

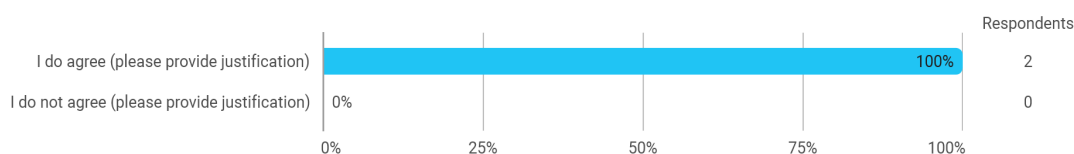
In addition, the experts had the opportunity (optional) to express their agreement or disagreement with individual update suggestions, including justifications. This feedback is presented in Appendices A, B and C. As expected, a wide variety of comments were obtained. Sometimes all responding experts agreed on the update suggestion, while in other cases all disagreed. Often the opinions were different so that no final conclusion could be drawn.

The variety of comments is demonstrated in the following examples of the feedback on single update suggestions.

Example OECD TG 301

4. para 2 - Acknowledgment that the proposed methods are not equivocal in their performance and their outcome is required.

Rationale: There are seven different types of test designs, which are widely used, and inocula can be selected from five different environmental sources with different concentrations and microbial communities (OECD 1992; OECD 2014; Kowalczyk et al. 2015). The inocula can further be prepared in up to four different ways (Goodhead et al., 2014). Consequently, RBT outcomes for a given substance can differ widely (Thouand et al. 1995; Vazquez-Rodriguez et al. 2003; Goodhead et al. 2014). This does imply that an unequivocal result is achievable.



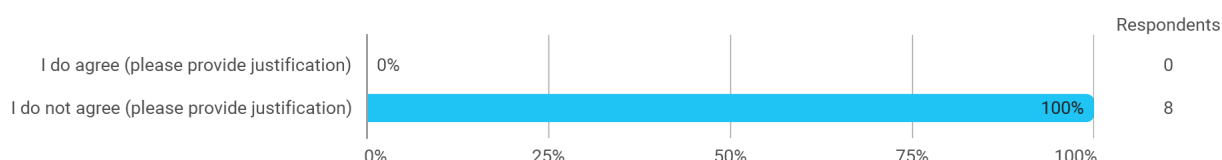
Agreement justifications

- A brief acknowledgement would be useful

- Propose to include the rationale along with this suggestion, as follows: There are seven different types of test designs, which are widely used, and inocula can be selected from five different environmental sources with different concentrations and microbial communities (OECD 1992; OECD 2014; Kowalczyk et al. 2015). The inocula can further be prepared in up to four different ways (Goodhead et al., 2014). Consequently, RBT outcomes for a given substance can differ widely (Thouand et al. 1995; Vazquez-Rodriguez et al. 2003; Goodhead et al. 2014). This does imply that an unequivocal result is achievable.

Example OECD TG 309

- para 16, lines 3-4: Tests should be run at least in triplicate because intra study variability has been observed to be high, particularly for low or no sediment (Seller et al., 2020)**



Disagreement justifications

- Even with 3 variability could occur - should than be done with each degradation simulation OECD test
- See previous. Duplicate is sufficient for ^{14}C studies
- The number of replicates is sufficient when radiolabelled test substance is used. Additionally, since a time series of data points is analysed, outliers are easily identified and can be addressed
- more replicates should be optional; for ^{14}C -labelled test items, 2 replicates are sufficient for detecting outliers
- Depends on the substance, for most substances we observed low variability between replicates. To improve kinetic calculations, it is advisable to increase the number of samplings instead of number of replicates
- Seller et al. (2020) investigated water-sediment systems only.
- A larger number of replicates may be useful when applying a non-labelled test item. In case of a radio-labelled test item, the analytical variability is lower, and 2 replicates should be sufficient. Also, degradation is measured at several sampling intervals and therefore a chance to identify outliers is given.

As providing justifications on agreement or disagreement with single update suggestions was optional, not all experts, who gave feedback on the general relevance for a specific TG revision provided feedback on all individual update suggestions for that TG. Therefore, the number of feedbacks was sometimes limited. In the first example above only two experts provided specific feedback, but 27 replied on the general relevance of the TG revision. In contrast, in the second example eight of 27 experts disagreed on the suggested update. Therefore, it is expected that the results obtained need to be used with caution. Even if a majority of respondents agreed or disagreed, the provided update suggestion may still be plausible, as not all background information might have been considered during the pre-evaluation process. Moreover, the

suggested update might have been misunderstood by the experts because the rationales were not provided and an in-depth discussion within the community would have been necessary. For these reasons, the specific statements for each single suggestion might not be very robust. Nevertheless, the feedback will provide a valuable basis for further discussion within potential upcoming OECD Expert Groups. They can be found in the Appendices A, B and C to this report.

5 Key results – Relevance and Prioritisation

During the pre-evaluation survey stakeholders were asked to provide a rating on the general relevance of TG revision based on the submitted update suggestions. The relevance of the proposed revisions of the different TGs was then assessed by the project team in three different ways. First, the TGs were ranked according to the number of update suggestions made during the online survey. Secondly, those TGs were identified that were rated as high by the majority of participants in the pre-evaluation in terms of relevance of revision (expressed in %). Finally, the rating of the individual TGs (low/medium/high relevance of revision) was further refined by weighting the results obtained. For this purpose, the specific rating for each individual TG was weighted by multiplying the percentages obtained for low, medium and high relevance of revision with factors 1, 2 and 3, respectively. The weighted percentage value was then calculated according to the following formula [weighted % = (%low x1 + %medium x3 + %high x5)/500]. Based on the results obtained by the different ranking approaches, the relevance of revision of the individual guidelines was prioritised.

5.1 Block “Fate”

5.1.1 Online survey: Number of update suggestions – Top 5 “Fate”

With 55 submissions, OECD TG 309 received by far the most update suggestions during the online survey (Block “Fate”) followed by OECD TG 301 and 308. The TOP 5 is complemented by TG 307 and 106.

1. Test No. 309: Aerobic Mineralisation in Surface Water – Simulation Biodegradation Test (55 update suggestions)
2. Test No. 301: Ready Biodegradability (36 update suggestions)
3. Test No. 308: Aerobic and Anaerobic Transformation in Aquatic Sediment Systems (34 update suggestions)
4. Test No. 307: Aerobic and Anaerobic Transformation in Soil (28 update suggestions)
5. Test No. 106: Adsorption -- Desorption Using a Batch Equilibrium Method (13 update suggestions)

5.1.2 Pre-evaluation of suggestions – relevance of TG revision “Fate”

The relevance of TG revision (low/medium/high relevance of revision in %) is presented in Table 2. With 81%, the highest percentage for “high relevance of revision” was expressed for OECD TG 301 followed by OECD TG 309 and 305 with 70%, respectively. Only a low or medium relevance of revision was identified for OECD TG 113 and 312. It must be noted that the individual assessment of the TGs was based on different numbers of replies ranging from 5 to 28 per TG.

Table 2: Relevance of TG revision (%) in Block “Fate” ranked according to “high relevance”

OECD TG	Total replies	Low [%]	Medium [%]	High [%]
Test No. 301: Ready Biodegradability	27	7	11	81
Test No. 309: Aerobic Mineralisation in Surface Water – Simulation Biodegradation Test	27	22	7	70
Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure	10	0	30	70
Test No. 310: Ready Biodegradability – CO ₂ in sealed vessels (Headspace Test)	25	12	20	68
OECD Draft Soil Photolysis guideline (2002)	18	22	11	67
Test No. 306: Biodegradability in Seawater	23	22	17	61
Test No. 107: Partition Coefficient (n-octanol/water): Shake Flask Method	15	7	33	60
Test No. 308: Aerobic and Anaerobic Transformation in Aquatic Sediment Systems	26	19	23	58
Test No. 307: Aerobic and Anaerobic Transformation in Soil	28	21	21	57
Test No. 302B: Inherent Biodegradability: Zahn-Wellens/ EVPA Test	24	21	25	54
Test No. 105 – Water solubility	24	4	46	50
Test No. 106: Adsorption – Desorption Using a Batch Equilibrium Method	18	17	39	44
Test No. 117: Partition Coefficient (n-octanol/water), HPLC Method	15	13	47	40
Test No. 111: Hydrolysis as a Function of pH	15	40	20	40
Test No. 121: Estimation of the Adsorption Coefficient (K _{oc}) on Soil and on Sewage Sludge using High Performance Liquid Chromatography (HPLC)	13	23	38	38
Test No. 115: Surface Tension of Aqueous Solutions	8	25	38	38
Revised Introduction to the OECD Guidelines for Testing of Chemicals, Section 3	28	7	57	36
Test No. 122: Determination of pH, Acidity and Alkalinity	9	33	33	33
Test No. 316: Phototransformation of Chemicals in Water – Direct Photolysis	18	22	50	28
Test No. 319B: Determination of in vitro intrinsic clearance using rainbow trout liver S9 sub-cellular fraction (RT-S9)	5	20	60	20
Test No. 123: Partition Coefficient (1-Octanol/Water): Slow-Stirring Method	11	18	64	18

OECD TG	Total replies	Low [%]	Medium [%]	High [%]
Test No. 112: Dissociation Constants in Water	12	8	75	17
Test No. 318: Dispersion Stability of Nanomaterials in Simulated Environmental Media	14	50	36	14
Test No. 113: Screening Test for Thermal Stability and Stability in Air	10	40	60	0
Test No. 312: Leaching in Soil Columns	10	50	50	0
Test 209: Activated Sludge, Respiration Inhibition Test (Carbon and Ammonium Oxidation)	10	20	70	10

When assessed on the basis of weighted % values, OECD TG 301 was also considered the most relevant one for a potential revision. For four TGs a weighted % value higher than 80 was estimated indicating a particularly high relevance of revision. With weighted % values of 46, 44, and 40, the lowest relevance of TG revision was estimated for OECD TGs 318, 113 and 312, respectively.

Table 3: Relevance of TG revision in Block “Fate” ranked according to weighted relevance [%]

OECD TG	Total replies	Low [%]	Medium [%]	High [%]	Weighted* [%]
Test No. 301: Ready Biodegradability	27	7	11	81	90
Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure	10	0	30	70	88
Test No. 310: Ready Biodegradability - CO ₂ in sealed vessels (Headspace Test)	25	12	20	68	82
Test No. 107: Partition Coefficient (n-octanol/water): Shake Flask Method	15	7	33	60	81
Test No. 309: Aerobic Mineralisation in Surface Water – Simulation Biodegradation Test	27	22	7	70	79
Test No. 105 - Water solubility	24	4	46	50	78
OECD Draft Soil Photolysis guideline (2002)	18	22	11	67	78
Test No. 306: Biodegradability in Seawater	23	22	17	61	76
Test No. 308: Aerobic and Anaerobic Transformation in Aquatic Sediment Systems	26	19	23	58	75
Test No. 307: Aerobic and Anaerobic Transformation in Soil	28	21	21	57	74
Test No. 302B: Inherent Biodegradability: Zahn-Wellens/ EVPA Test	24	21	25	54	73
Revised Introduction to the OECD Guidelines for Testing of Chemicals, Section 3	28	7	57	36	71

OECD TG	Total replies	Low [%]	Medium [%]	High [%]	Weighted* [%]
Test No. 106: Adsorption -- Desorption Using a Batch Equilibrium Method	18	17	39	44	71
Test No. 117: Partition Coefficient (n-octanol/water), HPLC Method	15	13	47	40	71
Test No. 121: Estimation of the Adsorption Coefficient (Koc) on Soil and on Sewage Sludge using High Performance Liquid Chromatography (HPLC)	13	23	38	38	66
Test No. 115: Surface Tension of Aqueous Solutions	8	25	38	38	65
Test No. 112: Dissociation Constants in Water	12	8	75	17	63
Test No. 316: Phototransformation of Chemicals in Water – Direct Photolysis	18	22	50	28	62
Test No. 123: Partition Coefficient (1-Octanol/Water): Slow-Stirring Method	11	18	64	18	60
Test No. 319B: Determination of in vitro intrinsic clearance using rainbow trout liver S9 sub-cellular fraction (RT-S9)	5	20	60	20	60
Test No. 122: Determination of pH, Acidity and Alkalinity	9	33	33	33	60
Test No. 111: Hydrolysis as a Function of pH	15	40	20	40	60
Test 209: Activated Sludge, Respiration Inhibition Test (Carbon and Ammonium Oxidation)	10	20	70	10	46
Test No. 318: Dispersion Stability of Nanomaterials in Simulated Environmental Media	14	50	36	14	46
Test No. 113: Screening Test for Thermal Stability and Stability in Air	10	40	60	0	44
Test No. 312: Leaching in Soil Columns	10	50	50	0	40

*) weighted % = [(%low x1 + %medium x3 + %high x5)/500]

Pre-evaluation of update suggestions – Relevance of TG revision (Top 3 – Lowest 4)

1. Test No. 301: Ready Biodegradability
2. Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure
3. Test No. 310: ready Biodegradability – CO2 in sealed vessels (Headspace test)
- ...
24. Test 209: Activated Sludge, Respiration Inhibition Test (Carbon and Ammonium Oxidation)
25. Test No. 318: Dispersion Stability of Nanomaterials in Simulated Environmental Media
26. Test No. 113: Screening Test for Thermal Stability and Stability in Air
27. Test No. 312: Leaching in Soil Columns

5.2 Block “Terrestrial”

5.2.1 Online survey: Number of update suggestions – Top 5 “Terrestrial”

During the online survey the Top 5 OECD TGs of Block B „Terrestrial“ obtained 36 to 24 update suggestions with OECD TG 206 being at the top of the list, followed by TG 227, 213, 222 and 208.

1. Test No. 206: Avian Reproduction Test (36 suggestions)
2. Test No. 227: Terrestrial Plant Test: Vegetative Vigour Test (33 suggestions)
3. Test No. 213: Honeybees, Acute Oral Toxicity Test (29 suggestions)
4. Test No. 222: Earthworm Reproduction Test (*Eisenia fetida*/*Eisenia andrei*) (26 suggestions)
5. Test No. 208: Terrestrial Plant Test: Seedling Emergence and Seedling Growth Test (24 suggestions)

5.2.2 Pre-evaluation of suggestions – relevance of TG revision “Terrestrial”

Regarding the relevance of TG revision (%), a high proportion of participants (>75%) rated the revision need of OECD TGs 208, 227, and 239 as high. However, it must be noted that not all TGs were evaluated by the same number of participants (Table 4).

Table 4: Relevance of TG revision (%) in Block “Terrestrial” ranked according to “high relevance”

OECD TG	Total replies	Low [%]	Medium [%]	High [%]
Test No. 208: Terrestrial Plant Test: Seedling Emergence and Seedling Growth Test	21	0	14	86
Test No. 227: Terrestrial Plant Test: Vegetative Vigour Test	16	0	19	81
Guidance Document 239: Honey Bee Larval	24	13	13	75
Test No. 222: Earthworm Reproduction Test (<i>Eisenia fetida</i> / <i>Eisenia andrei</i>)	22	5	23	73
Test No. 206: Avian Reproduction Test	7	14	14	71
Test No. 214: Honeybees, Acute Contact Toxicity Test	22	0	32	68
Test No. 226: Predatory mite (<i>Hypoaspis</i> (<i>Geolaelaps</i>) <i>aculeifer</i>) reproduction test in soil	22	5	27	68
Test No. 213: Honeybees, Acute Oral Toxicity Test	25	0	32	68
Test No. 232: Collembolan Reproduction Test in Soil	20	15	20	65
Test No. 223: Avian Acute Oral Toxicity Test	7	14	29	57
Test No. 216: Soil Microorganisms: Nitrogen Transformation Test	23	0	48	52
Test No. 237: Honey Bee (<i>Apis mellifera</i>) Larval Toxicity Test, Single Exposure	25	20	28	52
Test No. 207: Earthworm, Acute Toxicity Tests	29	24	34	41
Test No. 246: Bumblebee, Acute Contact Toxicity Test	21	24	38	38

OECD TG	Total replies	Low [%]	Medium [%]	High [%]
Test No. 245: Honey Bee (<i>Apis mellifera</i> L.), Chronic Oral Toxicity Test (10-Day Feeding)	21	33	33	33
Test No. 205: Avian Dietary Toxicity Test	16	56	19	25
Test No. 247: Bumblebee, Acute Oral Toxicity Test	21	29	52	19
Test No. 220: Enchytraeid Reproduction Test	18	44	39	17
Test No. 228: Determination of Developmental Toxicity to Dipteran Dung Flies	9	44	56	0

A similar result was obtained when the assessment was based on weighted % values (Table 5), with OECD TGs 208 and 227 again being considered as the TGs with the highest relevance for TG revision. In the terrestrial block, a weighted % value greater 80 was estimated for nine TGs, indicating a broader relevance of revision. However, also in this case TGs with weighted % values below 50 indicating a lower relevance of TG revision were identified, e.g., OECD TGs 220, 205 and 228.

Table 5: Relevance of TG revision in Block “Terrestrial” ranked according to weighted relevance [%]

OECD TG	Total replies	Low [%]	Medium [%]	High [%]	Weighted* [%]
Test No. 208: Terrestrial Plant Test: Seedling Emergence and Seedling Growth Test	21	0	14	86	94
Test No. 227: Terrestrial Plant Test: Vegetative Vigour Test	16	0	19	81	93
Test No. 222: Earthworm Reproduction Test (<i>Eisenia fetida</i> / <i>Eisenia andrei</i>)	22	5	23	73	87
Test No. 214: Honeybees, Acute Contact Toxicity Test	22	0	32	68	87
Test No. 213: Honeybees, Acute Oral Toxicity Test	25	0	32	68	87
Test No. 226: Predatory mite (<i>Hypoaspis</i> (<i>Geolaelaps</i>) <i>aculeifer</i>) reproduction test in soil	22	5	27	68	85
Guidance Document 239: Honey Bee Larval	24	13	13	75	85
Test No. 206: Avian Reproduction Test	7	14	14	71	83
Test No. 216: Soil Microorganisms: Nitrogen Transformation Test	23	0	48	52	81
Test No. 232: Collembolan Reproduction Test in Soil	20	15	20	65	80
Test No. 223: Avian Acute Oral Toxicity Test	7	14	29	57	77
Test No. 237: Honey Bee (<i>Apis mellifera</i>) Larval Toxicity Test, Single Exposure	25	20	28	52	73
Test No. 207: Earthworm, Acute Toxicity Tests	29	24	34	41	67

OECD TG	Total replies	Low [%]	Medium [%]	High [%]	Weighted* [%]
Test No. 246: Bumblebee, Acute Contact Toxicity Test	21	24	38	38	66
Test No. 245: Honey Bee (<i>Apis mellifera</i> L.), Chronic Oral Toxicity Test (10-Day Feeding)	21	33	33	33	60
Test No. 247: Bumblebee, Acute Oral Toxicity Test	21	29	52	19	56
Test No. 220: Enchytraeid Reproduction Test	18	44	39	17	49
Test No. 205: Avian Dietary Toxicity Test	16	56	19	25	48
Test No. 228: Determination of Developmental Toxicity to Dipteran Dung Flies	9	44	56	0	42

*) weighted % = $[(\%low \times 1 + \%medium \times 3 + \%high \times 5)/500]$

Pre-evaluation of suggestions –Relevance of TG revision (Top 3 – Lowest 3)

1. Test No. 208: Terrestrial Plant Test: Seedling Emergence and Seedling Growth Test
2. Test No. 227: Terrestrial Plant Test: Vegetative Vigour Test
3. Test No. 222: Earthworm Reproduction Test (*Eisenia fetida*/*Eisenia andrei*)
- ...
17. Test No. 220: Enchytraeid Reproduction Test
18. Test No. 205: Avian Dietary Toxicity Test
19. Test No. 228: Determination of Developmental Toxicity to Dipteran Dung Flies

5.3 Block “Aquatic”

5.3.1 Online survey: Number of update suggestions – Top 5 “Aquatic”

During the online survey for no OECD TG being part of the block “Aquatic” more than 24 update suggestions were received.

1. Test No. 201: Freshwater Alga and Cyanobacteria, Growth Inhibition Test (24 suggestions)
2. Test No. 211: Daphnia magna Reproduction Test (22 suggestions)
3. Test No. 202: Daphnia sp. Acute Immobilisation Test (18 suggestions)
4. Test No. 231: Amphibian Metamorphosis Assay (16 suggestions)
5. Test No. 234: Fish Sexual Development Test (15 suggestions)

5.3.2 Pre-evaluation of suggestions – relevance of TG revision “Aquatic”

During the pre-evaluation of the suggestions, for most of the TGs less than 50% of the evaluators rated the relevance of a TG revision need as high. However, with 78% and 76%, a high relevance of TG revision was identified for OECD TG 231 and 201, respectively (Table 6).

Table 6: Relevance of TG revision (%) in Block “Aquatic” ranked according to “high relevance”

OECD TG	Total replies	Low [%]	Medium [%]	High [%]
Test No. 231: Amphibian Metamorphosis Assay	18	0	22	78
Test No. 201: Freshwater Alga and Cyanobacteria, Growth Inhibition Test	21	0	24	76
Test No. 229: Fish Short Term Reproduction Assay	24	8	29	63
Test No. 230: 21-day Fish Assay	22	14	32	55
Test No. 234: Fish Sexual Development Test	22	0	45	55
Test No. 241: The Larval Amphibian Growth and Development Assay (LAGDA)	16	13	38	50
Test No. 212: Fish, Short-term Toxicity Test on Embryo and Sac-Fry Stages	22	5	45	50
Test No. 210: Fish, Early-life Stage Toxicity Test	27	11	44	44
Test No. 248: Xenopus Eleutheroembryonic Thyroid Assay (XETA)	14	14	43	43
Test No. 240: Medaka Extended One Generation Reproduction Test (MEOGRT)	17	18	41	41
Test No. 236: Fish Embryo Acute Toxicity (FET) Test	23	9	52	39
Test No. 211: Daphnia magna Reproduction Test	18	0	61	39
Test No. 203: Fish, Acute Toxicity Test	35	9	54	37
Test No. 239: Water-Sediment Myriophyllum Spicatum Toxicity Test	17	12	53	35
Test No. 202: Daphnia sp. Acute Immobilisation Test	22	9	59	32
Test No. 235: Chironomus sp., Acute Immobilisation Test	13	23	46	31
Test No. 215: Fish, Juvenile Growth Test	21	10	62	29
Test No. 221: Lemna sp. Growth Inhibition Test	19	16	58	26
Test No. 238: Sediment-Free Myriophyllum Spicatum Toxicity Test	18	11	67	22
Test No. 218: Sediment-Water Chironomid Toxicity Using Spiked Sediment	14	7	71	21
Test No. 219: Sediment-Water Chironomid Toxicity Using Spiked Water	14	7	71	21
Test 209: Activated Sludge, Respiration Inhibition Test (Carbon and Ammonium Oxidation)	10	20	70	10
Test No. 243: <i>Lymnaea stagnalis</i> Reproduction Test	10	40	50	10

OECD TG	Total replies	Low [%]	Medium [%]	High [%]
Test No. 225: Sediment-Water Lumbriculus Toxicity Test Using Spiked Sediment	10	30	70	0
Test No. 233: Sediment-Water Chironomid Life-Cycle Toxicity Test Using Spiked Water or Spiked Sediment	11	27	73	0

Ranking the TGs in terms of relevance of TG revision based on weighted % values, OECD TG 231, 201, and 234, ranked top 3 on the list. The lowest relevance for TG revision was estimated for OECD TG 233, 209, 243, and 225 with weighted % values below 50 (Table 7).

Table 7: Relevance of TG revision in Block “Aquatic” ranked according to weighted relevance [%]

OECD TG	Total replies	Low [%]	Medium [%]	High [%]	Weighted* [%]
Test No. 231: Amphibian Metamorphosis Assay	18	0	22	78	91
Test No. 201: Freshwater Alga and Cyanobacteria, Growth Inhibition Test	21	0	24	76	90
Test No. 234: Fish Sexual Development Test	22	0	45	55	82
Test No. 229: Fish Short Term Reproduction Assay	24	8	29	63	82
Test No. 212: Fish, Short-term Toxicity Test on Embryo and Sac-Fry Stages	22	5	45	50	78
Test No. 230: 21-day Fish Assay	22	14	32	55	76
Test No. 211: Daphnia magna Reproduction Test	18	0	61	39	76
Test No. 241: The Larval Amphibian Growth and Development Assay (LAGDA)	16	13	38	50	75
Test No. 210: Fish, Early-life Stage Toxicity Test	27	11	44	44	73
Test No. 236: Fish Embryo Acute Toxicity (FET) Test	23	9	52	39	72
Test No. 203: Fish, Acute Toxicity Test	35	9	54	37	71
Test No. 248: Xenopus Eleutheroembryonic Thyroid Assay (XETA)	14	14	43	43	71
Test No. 240: Medaka Extended One Generation Reproduction Test (MEOGRT)	17	18	41	41	69
Test No. 239: Water-Sediment Myriophyllum Spicatum Toxicity Test	17	12	53	35	69
Test No. 202: Daphnia sp. Acute Immobilisation Test	22	9	59	32	69
Test No. 215: Fish, Juvenile Growth Test	21	10	62	29	68
Test No. 218: Sediment-Water Chironomid Toxicity Using Spiked Sediment	14	7	71	21	66

OECD TG	Total replies	Low [%]	Medium [%]	High [%]	Weighted*
Test No. 219: Sediment-Water Chironomid Toxicity Using Spiked Water	14	7	71	21	66
Test No. 238: Sediment-Free Myriophyllum Spicatum Toxicity Test	18	11	67	22	64
Test No. 221: Lemna sp. Growth Inhibition Test	19	16	58	26	64
Test No. 235: Chironomus sp., Acute Immobilisation Test	13	23	46	31	63
Test No. 233: Sediment-Water Chironomid Life-Cycle Toxicity Test Using Spiked Water or Spiked Sediment	11	27	73	0	49
Test No. 243: <i>Lymnaea stagnalis</i> Reproduction Test	10	40	50	10	48
Test No. 225: Sediment-Water Lumbriculus Toxicity Test Using Spiked Sediment	10	30	70	0	48
Test 209: Activated Sludge, Respiration Inhibition Test (Carbon and Ammonium Oxidation)	10	20	70	10	46

*) weighted % = [(%low x1 + %medium x3 + %high x5)/500]

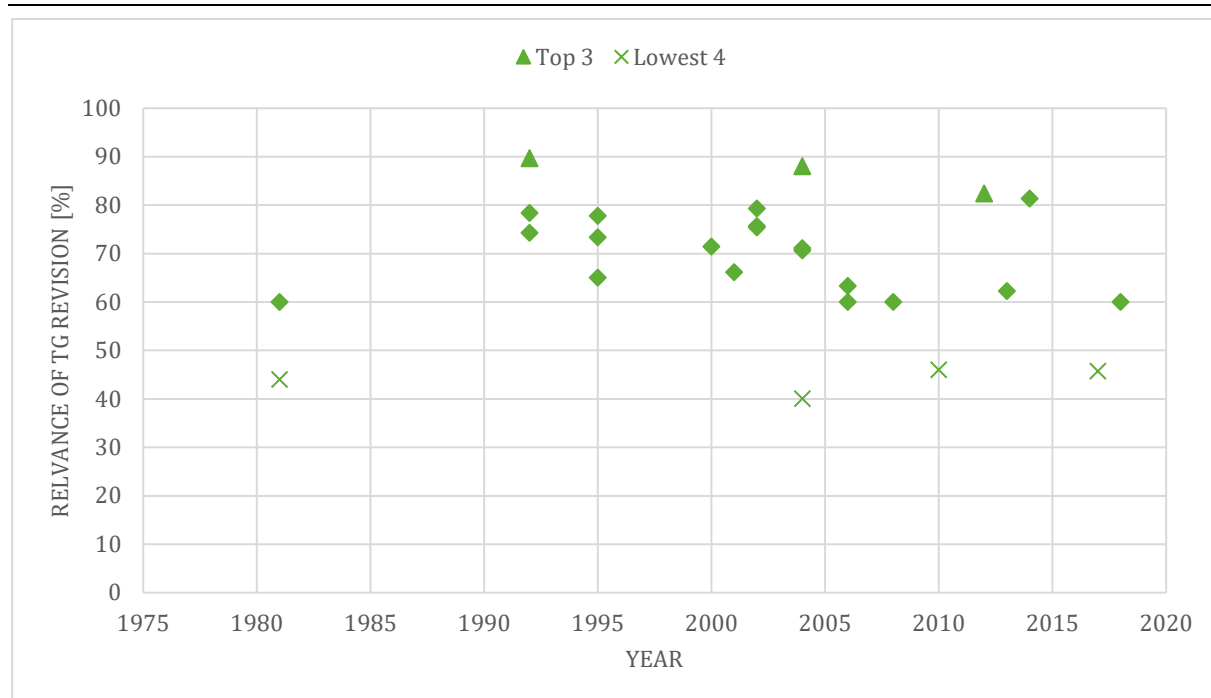
Pre-evaluation of suggestions – Relevance of TG revision (Top 3 – Lowest 4)

1. Test No. 231: Amphibian Metamorphosis Assay
2. Test No. 201: Freshwater Alga and Cyanobacteria, Growth Inhibition Test
3. Test No. 234: Fish Sexual Development Test
- ...
23. Test No. 243: *Lymnaea stagnalis* Reproduction Test
24. Test No. 225: Sediment-Water Lumbriculus Toxicity Test Using Spiked Sediment
25. Test No. 209 Activated Sludge, Respiration Inhibition Test (Carbon and Ammonium Oxidation)

5.4 Relevance of year of publication

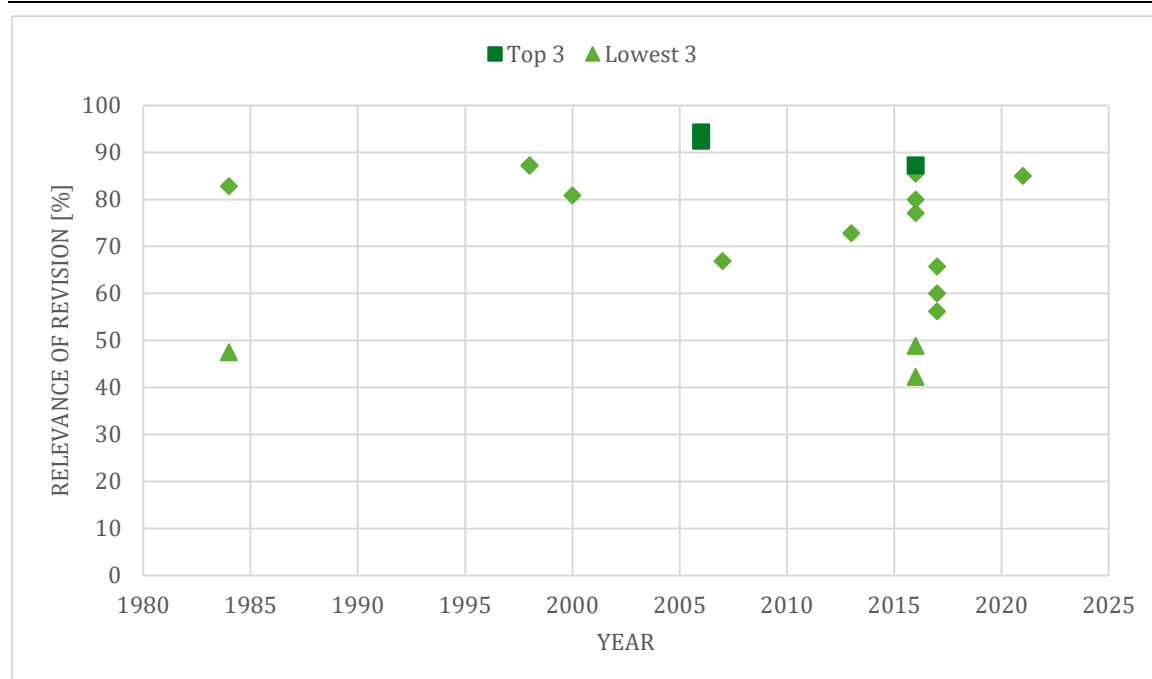
In a further approach, the relevance of TG revision (weighted %) estimated for each guideline was compared with the TG's year of publication. The results for block "fate", "terrestrial" and "aquatic" are presented in Figures 6-8. The results show that in all blocks no correlation was found between both factors even if only the Top 3 and Lowest 3 TGs of the rankings were considered.

Figure 6: Pre-evaluation of suggestions – Relevance of TG revision (weighted %) vs. Year of publication; Block “Fate”



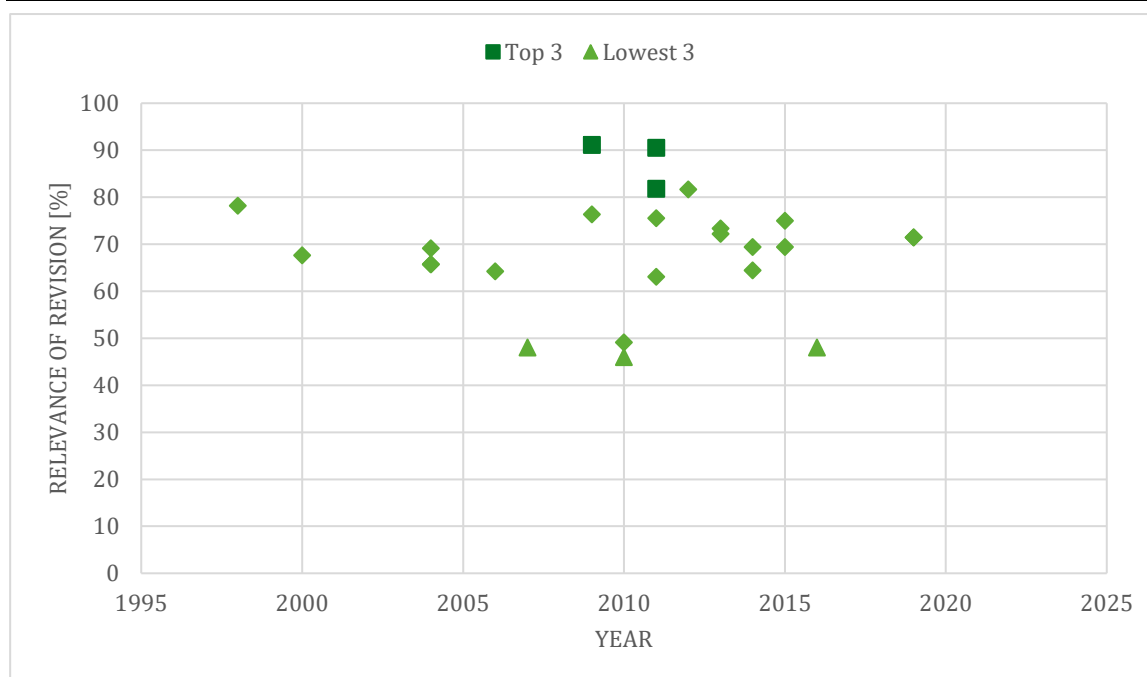
Source: own illustration, Fraunhofer Institute

Figure 7: Pre-evaluation of suggestions – Relevance of TG revision (weighted %) vs. Year of publication; Block “Terrestrial”



Source: own illustration, Fraunhofer Institute

Figure 8: Pre-evaluation of suggestions – Relevance of TG revision (weighted %) vs. Year of publication; Block “Aquatic”



Source: own illustration, Fraunhofer Institute

5.5 Interest in OECD Expert Group participation

Participants of the pre-evaluation survey were asked about their interest in participating in a potential OECD Expert Group dealing with the revision of a specific guideline. The results are compared with the relevance of revision estimated for the different TGs (Figures 9-11). Interestingly, TGs which were identified to have the highest relevance of revision have not necessarily generated the greatest interest in an OECD Expert Group participation (Figure 9). This might be at least partly explained by the limited circle of experts being involved in the usage of very specific TGs as shown for TG 223 and TG 206 (Figure 10). On the other hand, TGs can be of interest to a wide group of stakeholders, leading to a high interest in OECD Expert Group participation as shown for TG 203 and TG 210 (Figure 11).

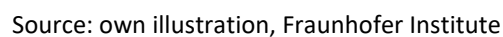
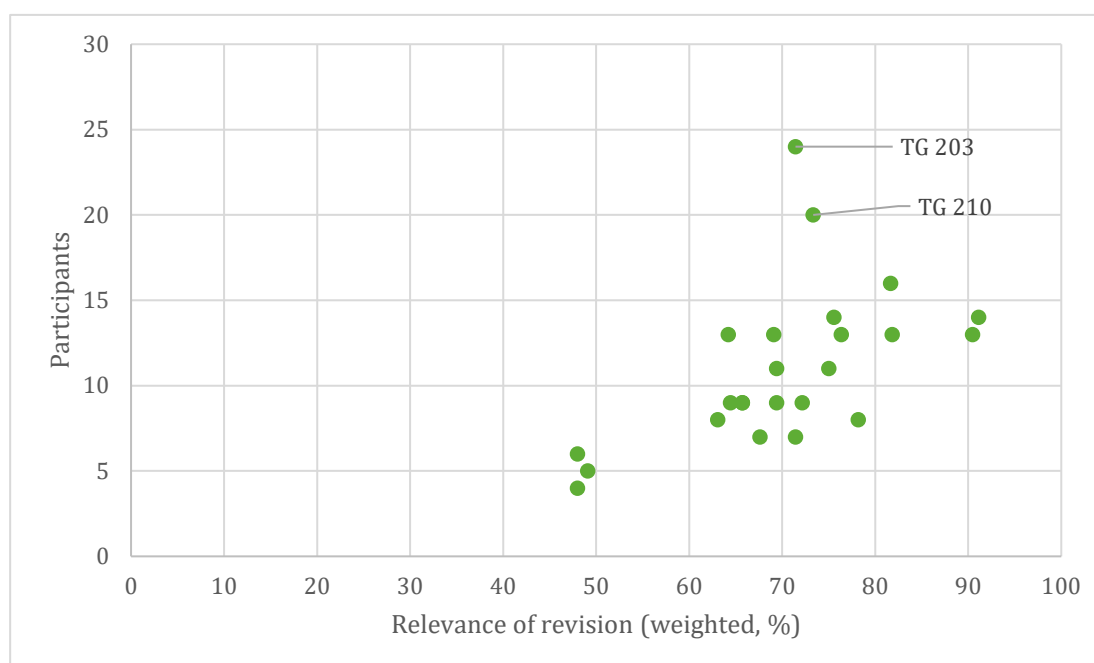


Figure 11: Interest in OECD Expert Group participation (Block “Aquatic”)



Source: own illustration, Fraunhofer Institute

5.6 Conclusions on the online and pre-evaluation survey

The project was intended to review which OECD TGs do not correspond to the current state of science and technology. All TGs relevant for the assessment of effects of chemicals on the environment were reviewed. The following conclusions can be drawn:

- ▶ High number of suggestions to further improve OECD TGs were received (max 55 per TG)
- ▶ Evaluation of relevance of TG revision: High number of respondents per TG (5-35)
- ▶ Pre-evaluation of suggestions: Manifold comments which express agreement/disagreement with update suggestions and provide thorough basis for further discussion
- ▶ TGs with high relevance of revision were identified
- ▶ Relevance of revision is independent of TGs' year of publication / last revision
- ▶ High interest in OECD Expert Group participation

5.7 International workshops

The results of the project were discussed with all interested parties including industry, science, contract labs, NGOs and authorities. Three virtual workshops were conducted covering one of the three different blocks (“fate”, “terrestrial” and “aquatic”) each. Each workshop was offered on two different days and time slots. Important aspects that came out of the survey and pre-evaluation were discussed and enough time for the participants to report on their experiences in conducting the survey was foreseen. Several introductory questions were addressed to participants to lead the discussion which is summarized below:

Any improvement suggestions regarding the evaluation process? Are there any important points related to evaluation that were missed?

The survey was promoted in different ways by contacting the stakeholders via meetings (e.g., SETAC), regulatory committees, and associations. A total of 68 experts participated in the survey. Even if the first survey was successful with a great response resulting in 257 completed questionnaires including roughly 500 individual update suggestions, for a follow-up project the dissemination of information should be further improved to reach a wider circle of experts. A higher participation of experts from non-EU countries, academia and NGOs would be desirable to achieve a more balanced field of participants.

Contribution of experts from regions around the world was under representative for non-EU countries and some stakeholder groups as science and NGOs (see also Fig 1 and 2 in chapter 3.1.2). Experts from industry, science, contract labs, NGOs and authorities were among the participants of the survey, however, for several TGs suggestions were only received by part of the stakeholders. The received suggestions therefore do not necessarily reflect the opinion of the entire field of participants.

The experience gained from the survey and pre-validation process should feed into future TG evaluation projects. Above all, consideration should be given to A) an extended commenting period to allow careful revision of the large set of TGs, B) provision of an improved website entry page with additional features like an option for intermediate caching of incomplete entries, as well as C) refined options for selecting documents for further processing. An attempt should also be made to consider the established processes applied in the evaluation of other directive systems (e.g., DIN/ISO). All this will help to further improve the evaluation process in the future.

Do you think a continuous TG evaluation process is helpful? If yes, in which time periods?

There was clear agreement that the evaluation process is helpful and should be continuously repeated. The majority of participants would welcome a repetition of the survey every 5 years. However, the length of the time required for updating the TGs should be also taken into account. Therefore, longer intervals for TG evaluation might be appropriate but which should not last longer than 10 to 15 years. Alternatively, it could be considered to establish a constantly available system for entering suggestions (“complaints box”) to make the TG evaluation process more flexible. A yearly report summarizing incoming suggestions could be published and information fed into the regular evaluation process.

TG evaluation and updating: Who do you think should be responsible for such a process?

There was a clear preference for OECD to initiate and coordinate the future evaluation process. Comments and suggestions should be shared on the OECD platform and brought to the attention of international organisations, associations and regulatory committees, e.g., CEFIC, ECHA, SETAC. An OECD contact point to coordinate the evaluation process would be welcomed by the participants. However, there was agreement that in the TG revision process action should remain with the member states with the National Co-ordinator as main contact person. It was suggested that a lead country that once brought a TG into the OECD process may also take responsibility for the further revision process. Expert Groups which were formerly involved in the development of a TG might be reconstituted to support the revision process. If required new Expert Group members should be nominated involving as many stakeholder groups as possible. The established OECD Test Guideline update process and the role of the National Co-ordinators were briefly summarized in the workshops (see Chapter 6).

How do you interpret the key results of the project?

There was agreement that the results of the survey reflect a high need to update several OECD TGs. However, the results of the survey and pre-evaluation can only indicate which TGs should

be primarily treated. It was critically noted that a high ranking of a TG may simply reflect the TGs' frequent application and use within several regulations compared to less regulatory used (more specialised) TGs. Therefore, further in-depth evaluation of the comments and suggestions (see Appendix A) by the users of the TGs was recommended to decide on the need for a further revision process. This is also required because several comments and suggestions were contradictory and need further clarification.

Generally, the results of the survey reflect the desire of experts from all stakeholder groups to be more engaged into the OECD TG development processes. The suggestions submitted during the survey provide a list of open questions which should be considered in future research activities to support the development of improved TGs.

It was not part of the pre-evaluation process to analyse the submitted suggestions with regard to their (regulatory) acceptance. It is therefore essential that National Co-ordinators critically monitor the further TG revision process(es).

How can the results be used in a TG update process?

The survey provided manifold suggestions to improve a broad range of OECD TGs. However, according to the current rules even minor adaptations (e.g., minimal changes to wording for better readability) cannot be easily implemented and would lead to a lengthy process. The official update process should therefore be reformed. The definition of two different pathways was suggested: one for „general revision“ and one for „minor adaptation“. However, even minor changes may have a high relevance for the produced data and therefore for the regulatory use. It needs to be clarified beforehand where in the regulatory process data from OECD TGs are used, how other TGs would be affected by a TG update and how it may affect the risk assessment. All stakeholders should be involved in the update process to ensure all aspects to be covered. However, update suggestions might be very specific and require in-depth discussion by experts. A list of experts who would like to join a TG revision process was created as part of the pre-evaluation process. In specific cases the respondents might be active in relevant OECD Expert Groups which would facilitate to start the update process.

National Co-ordinators play a central role in the TG update process and are the contact point for national experts intending to contribute. With respect to the vast number of suggestions obtained in this project, the complete revision of the TGs cannot be initiated by single National Co-ordinators and co-ordination on a higher level is required. A strategy on how OECD can deal with all the information derived from the survey and pre-evaluation process needs to be developed. It was suggested that this issue should be discussed during a future meeting of the National Co-ordinators involved in the OECD Test Guidelines Programme (WNT).

The survey provided several suggestions which not only support the improvement of TGs but also their harmonization. In this context, it was suggested that harmonization with other international guidelines (e.g., USEPA/ASTM) should be also considered for further development.

How can OECD Guidance Documents be addressed?

The need for the development of further OECD GDs was discussed. Such documents may help to address overarching issues such as the use of “difficult to test substances”. However, it was noted that developing GDs is laborious and GDs may also be outdated, e.g., “statistics”, and thus need revision. Generally, TGs need to be as clear as possible and the update of TGs should be prioritized.

Following the six workshops, two further presentations were given to the national co-ordinators involved in the OECD Test Guidelines Programme (WNT) to summarize the outcome of the project, and to lead an open discussion on the use of the results obtained for the future TG

revision process. There was agreement that there is a high demand for updating certain TGs and efforts should be undertaken to establish a regular process to check update needs and to support the required revision process.

6 The OECD Test Guideline update process

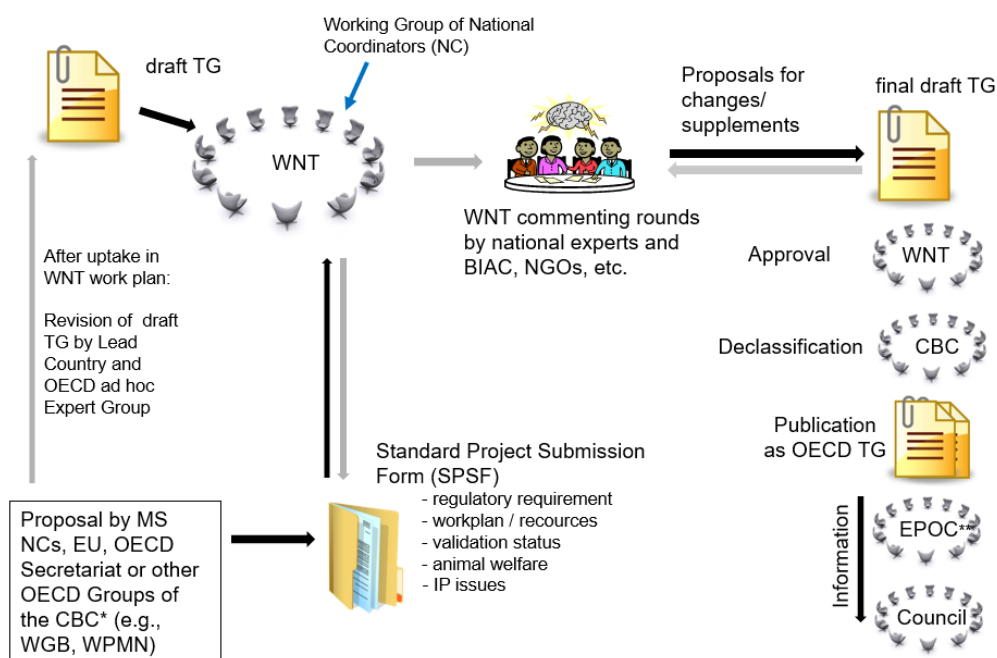
6.1 Official update process

The results of this project provide manifold starting points for the future OECD TG update process. There is an official process to follow for the development of new or the update of already existing TGs: (see OECD GD No. 1 2009; <https://www.oecd.org/chemicalsafety/testing/49803789.pdf>). The OECD Test Guideline Revision Process is presented in figure 12.

To remember:

- ▶ OECD TGs are internationally agreed standards
- ▶ Data are especially used for regulatory purpose worldwide
- ▶ OECD Council Decision of 1981: data from studies carried out in accordance with the OECD TGs and Principles of Good Laboratory Practice (GLP) in one OECD country must be accepted by other OECD countries for assessment purposes (Mutual Acceptance of Data)
- ▶ Any changes to OECD TGs are bound with workload for all member countries, OECD Secretariat and in particular the country/countries taking the lead for an update project

Figure 12: OECD Test Guideline Revision Process



Source: own illustration adapted from Schwirn, 2019 , UBA.

6.2 Ahead of the formal submission of a proposal

All stakeholders are encouraged to contribute to the development of new and revised OECD TGs. Ahead of the formal submission of a revision proposal the following aspects should be considered:

- ▶ Any expert/group of experts (industry, science, NGOs, etc.) that would like to propose an update, needs to get in contact with the respective National Co-ordinator designated for the OECD Test Guideline Programme (<https://www.oecd.org/chemicalsafety/testing/national-coordinators-test-guidelines-programme.htm>).
- ▶ Informal discussions with other member countries might take place related to potential interest in co-leading the project and availability of resources.
- ▶ The member country represented by the National Co-ordinator decides whether to take the lead (responsibility) for the update process of a certain OECD TG.

7 List of references

All TG Guidelines referred to in this report can be accessed via

<https://www.oecd.org/chemicalsafety/testing/oecdguidelinesforthetestingofchemicals.htm> (Q3, 2021)

OECD Guidance document on honeybee larval toxicity test following repeated exposure, Series on Testing and Assessment N. 239 (15th July 2016).

OECD (2019), Guidance Document on Aquatic Toxicity Testing of Difficult Substances and Mixtures, OECD Series on Testing and Assessment, OECD Publishing, Paris, <https://doi.org/10.1787/0ed2f88e-en>.

OECD (2005), Guidance Document on the validation and international acceptance of new and updated test methods for hazard assessment, OECD Series on Testing and Assessment, OECD Publishing, Paris

OECD (2006), Current approaches in the statistical analysis of ecotoxicity data: A guidance to application, OECD Series on Testing and Assessment, OECD Publishing, Paris

OECD (2006), Revised Introduction to the OECD Guidelines for Testing of Chemicals, Section 3, OECD Guidelines for the Testing of Chemicals, Section 3, OECD Publishing, Paris, <https://doi.org/10.1787/9789264030213-en>.

OECD (1995), Test No. 105: Water Solubility, OECD Guidelines for the Testing of Chemicals, Section 1, OECD Publishing, Paris, <https://doi.org/10.1787/9789264069589-en>.

OECD (2000), Test No. 106: Adsorption -- Desorption Using a Batch Equilibrium Method, OECD Guidelines for the Testing of Chemicals, Section 1, OECD Publishing, Paris, <https://doi.org/10.1787/9789264069602-en>.

OECD (1995), Test No. 107: Partition Coefficient (n-octanol/water): Shake Flask Method, OECD Guidelines for the Testing of Chemicals, Section 1, OECD Publishing, Paris, <https://doi.org/10.1787/9789264069626-en>.

OECD (2004), Test No. 111: Hydrolysis as a Function of pH, OECD Guidelines for the Testing of Chemicals, Section 1, OECD Publishing, Paris, <https://doi.org/10.1787/9789264069701-en>.

OECD (1981), Test No. 112: Dissociation Constants in Water, OECD Guidelines for the Testing of Chemicals, Section 1, OECD Publishing, Paris, <https://doi.org/10.1787/9789264069725-en>.

OECD (1981), Test No. 113: Screening Test for Thermal Stability and Stability in Air, OECD Guidelines for the Testing of Chemicals, Section 1, OECD Publishing, Paris, <https://doi.org/10.1787/9789264069749-en>.

OECD (1995), Test No. 115: Surface Tension of Aqueous Solutions, OECD Guidelines for the Testing of Chemicals, Section 1, OECD Publishing, Paris, <https://doi.org/10.1787/9789264069787-en>.

OECD (2022), Test No. 117: Partition Coefficient (n-octanol/water), HPLC Method, OECD Guidelines for the Testing of Chemicals, Section 1, OECD Publishing, Paris, <https://doi.org/10.1787/9789264069824-en>.

OECD (2001), Test No. 121: Estimation of the Adsorption Coefficient (K_{oc}) on Soil and on Sewage Sludge using High Performance Liquid Chromatography (HPLC), OECD Guidelines for the Testing of Chemicals, Section 1, OECD Publishing, Paris, <https://doi.org/10.1787/9789264069909-en>.

OECD (2013), Test No. 122: Determination of pH, Acidity and Alkalinity, OECD Guidelines for the Testing of Chemicals, Section 1, OECD Publishing, Paris, <https://doi.org/10.1787/9789264203686-en>.

OECD (2022), Test No. 123: Partition Coefficient (1-Octanol/Water): Slow-Stirring Method, OECD Guidelines for the Testing of Chemicals, Section 1, OECD Publishing, Paris, <https://doi.org/10.1787/9789264015845-en>.

OECD (2011), Test No. 201: Freshwater Alga and Cyanobacteria, Growth Inhibition Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264069923-en>.

OECD (2004), Test No. 202: Daphnia sp. Acute Immobilisation Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264069947-en>.

- OECD (2019), Test No. 203: Fish, Acute Toxicity Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264069961-en>.
- OECD (1984), Test No. 205: Avian Dietary Toxicity Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070004-en>.
- OECD (1984), Test No. 206: Avian Reproduction Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070028-en>.
- OECD (1984), Test No. 207: Earthworm, Acute Toxicity Tests, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070042-en>.
- OECD (2006), Test No. 208: Terrestrial Plant Test: Seedling Emergence and Seedling Growth Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070066-en>.
- OECD (2013), Test No. 210: Fish, Early-life Stage Toxicity Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264203785-en>.
- OECD (2012), Test No. 211: Daphnia magna Reproduction Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264185203-en>.
- OECD (1998), Test No. 212: Fish, Short-term Toxicity Test on Embryo and Sac-Fry Stages, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070141-en>.
- OECD (1998), Test No. 213: Honeybees, Acute Oral Toxicity Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070165-en>.
- OECD (1998), Test No. 214: Honeybees, Acute Contact Toxicity Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070189-en>.
- OECD (2000), Test No. 215: Fish, Juvenile Growth Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070202-en>.
- OECD (2000), Test No. 216: Soil Microorganisms: Nitrogen Transformation Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070226-en>.
- OECD (2004), Test No. 218: Sediment-Water Chironomid Toxicity Using Spiked Sediment, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070264-en>.
- OECD (2000), Test No. 217: Soil Microorganisms: Carbon Transformation Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070240-en>.
- OECD (2016), Test No. 220: Enchytraeid Reproduction Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264264472-en>.
- OECD (2006), Test No. 221: Lemna sp. Growth Inhibition Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264016194-en>.
- OECD (2016), Test No. 222: Earthworm Reproduction Test (*Eisenia fetida*/*Eisenia andrei*), OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264264496-en>.
- OECD (2016), Test No. 223: Avian Acute Oral Toxicity Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264264519-en>.
- OECD (2007), Test No. 225: Sediment-Water Lumbriculus Toxicity Test Using Spiked Sediment, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264067356-en>.

- OECD (2016), Test No. 226: Predatory mite (*Hypoaspis* (*Geolaelaps*) *aculeifer*) reproduction test in soil, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264264557-en>.
- OECD (2006), Test No. 227: Terrestrial Plant Test: Vegetative Vigour Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264067295-en>.
- OECD (2016), Test No. 228: Determination of Developmental Toxicity to Dipteran Dung Flies (*Scathophaga stercoraria* L. (*Scathophagidae*), *Musca autumnalis* De Geer (*Muscidae*)), OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264264571-en>.
- OECD (2012), Test No. 229: Fish Short Term Reproduction Assay, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264185265-en>.
- OECD (2009), Test No. 230: 21-day Fish Assay: A Short-Term Screening for Oestrogenic and Androgenic Activity, and Aromatase Inhibition, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264076228-en>.
- OECD (2009), Test No. 231: Amphibian Metamorphosis Assay, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264076242-en>.
- OECD (2016), Test No. 232: Collembolan Reproduction Test in Soil, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264264601-en>.
- OECD (2010), Test No. 233: Sediment-Water Chironomid Life-Cycle Toxicity Test Using Spiked Water or Spiked Sediment, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264090910-en>.
- OECD (2011), Test No. 234: Fish Sexual Development Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264122369-en>.
- OECD (2011), Test No. 235: *Chironomus* sp., Acute Immobilisation Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264122383-en>.
- OECD (2013), Test No. 236: Fish Embryo Acute Toxicity (FET) Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264203709-en>.
- OECD (2013), Test No. 237: Honey Bee (*Apis Mellifera*) Larval Toxicity Test, Single Exposure, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264203723-en>.
- OECD (2014), Test No. 238: Sediment-Free *Myriophyllum Spicatum* Toxicity Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264224131-en>.
- OECD (2014), Test No. 239: Water-Sediment *Myriophyllum Spicatum* Toxicity Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264224155-en>.
- OECD (2015), Test No. 240: Medaka Extended One Generation Reproduction Test (MEOGRT), OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264242258-en>.
- OECD (2015), Test No. 241: The Larval Amphibian Growth and Development Assay (LAGDA), OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264242340-en>.
- OECD (2016), Test No. 243: *Lymnaea stagnalis* Reproduction Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264264335-en>.
- OECD (2017), Test No. 245: Honey Bee (*Apis Mellifera* L.), Chronic Oral Toxicity Test (10-Day Feeding), OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264284081-en>.

OECD (2017), Test No. 246: Bumblebee, Acute Contact Toxicity Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264284104-en>.

OECD (2017), Test No. 247: Bumblebee, Acute Oral Toxicity Test, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/9789264284128-en>.

OECD (2019), Test No. 248: Xenopus Eleutheroembryonic Thyroid Assay (XETA), OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/a13f80ee-en>.

OECD (2021), Test No. 249: Fish Cell Line Acute Toxicity - The RTgill-W1 cell line assay, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/c66d5190-en>.

OECD (2021), Test No. 250: EASZY assay - Detection of Endocrine Active Substances, acting through estrogen receptors, using transgenic tg(cyp19a1b:GFP) Zebrafish embryos, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/0a39b48b-en>.

OECD (2022), Test No. 251: Rapid Androgen Disruption Activity Reporter (RADAR) assay, OECD Guidelines for the Testing of Chemicals, Section 2, OECD Publishing, Paris, <https://doi.org/10.1787/da264d82-en>.

OECD (1992), Test No. 301: Ready Biodegradability, OECD Guidelines for the Testing of Chemicals, Section 3, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070349-en>.

OECD (1992), Test No. 302B: Inherent Biodegradability: Zahn-Wellens/ EVPA Test, OECD Guidelines for the Testing of Chemicals, Section 3, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070387-en>.

OECD (2012), Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure, OECD Guidelines for the Testing of Chemicals, Section 3, OECD Publishing, Paris, <https://doi.org/10.1787/9789264185296-en>.

OECD (1992), Test No. 306: Biodegradability in Seawater, OECD Guidelines for the Testing of Chemicals, Section 3, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070486-en>.

OECD (2002), Test No. 307: Aerobic and Anaerobic Transformation in Soil, OECD Guidelines for the Testing of Chemicals, Section 3, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070509-en>.

OECD (2002), Test No. 308: Aerobic and Anaerobic Transformation in Aquatic Sediment Systems, OECD Guidelines for the Testing of Chemicals, Section 3, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070523-en>.

OECD (2004), Test No. 309: Aerobic Mineralisation in Surface Water – Simulation Biodegradation Test, OECD Guidelines for the Testing of Chemicals, Section 3, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070547-en>.

OECD (2014), Test No. 310: Ready Biodegradability - CO₂ in sealed vessels (Headspace Test), OECD Guidelines for the Testing of Chemicals, Section 3, OECD Publishing, Paris, <https://doi.org/10.1787/9789264224506-en>.

OECD (2004), Test No. 312: Leaching in Soil Columns, OECD Guidelines for the Testing of Chemicals, Section 3, OECD Publishing, Paris, <https://doi.org/10.1787/9789264070561-en>.

OECD (2008), Test No. 316: Phototransformation of Chemicals in Water – Direct Photolysis, OECD Guidelines for the Testing of Chemicals, Section 3, OECD Publishing, Paris, <https://doi.org/10.1787/9789264067585-en>.

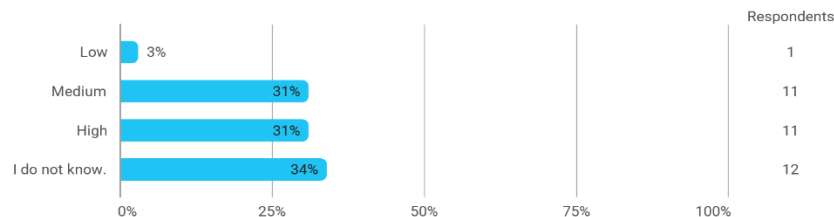
OECD (2017), Test No. 318: Dispersion Stability of Nanomaterials in Simulated Environmental Media, OECD Guidelines for the Testing of Chemicals, Section 3, OECD Publishing, Paris, <https://doi.org/10.1787/9789264284142-en>.

OECD (2018), Test No. 319B: Determination of in vitro intrinsic clearance using rainbow trout liver S9 sub-cellular fraction (RT-S9), OECD Guidelines for the Testing of Chemicals, Section 3, OECD Publishing, Paris, <https://doi.org/10.1787/9789264303232-en>.

A Assessment feedback Block A

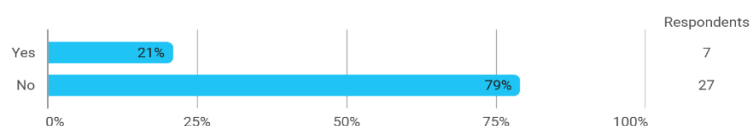
A.1 Test No. 105 - Water solubility

Relevance of revision



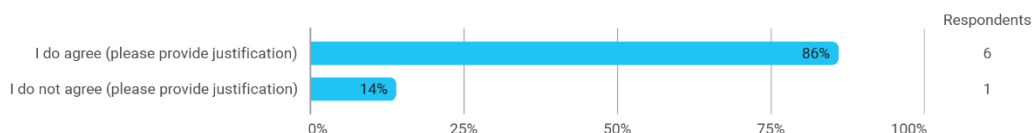
Interest in being part of an OECD

Expert Group



The following suggestions have been collected:

- 1. General: state that for sparingly soluble metals, information on water solubility should be generated using the OECD GD 29.**

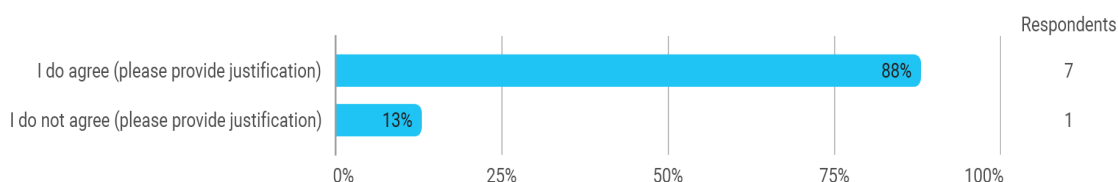


Agreement justifications:

- ▶ nice link
- ▶ It is important to have it also in this TG
- ▶ more information about the scope of application makes sense
- ▶ standard water solubility testing is not as informative for metals as the dissolution transformation test

- 2. para 21 - More recent references could be provided. New methods such as passive dosing could be addressed and other methods to address difficult to test substances. Could consider formalizing slow-stirring options into this guidance. Relevant reviews such as:**

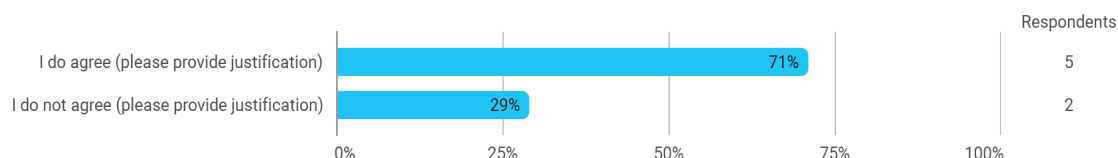
<https://www.sciencedirect.com/science/article/pii/S0003267019308426>



Agreement justifications

- ▶ relevant alternative methods
- ▶ Totally agree. There is a need for a complete revision of this TG for difficult to test substances
- ▶ considering further methods which became state of science in the meantime should be considered
- ▶ additional guidance for difficult substances are necessary
- ▶ There is a need to update this old TG to address difficult to test substances

3. Para 30 - Simply naming the analytical method may be inadequate. Further analytical details should be reported for confident evaluation and confirmation of substance identity, particularly moving towards more difficult substances: instrumentation detail for both separation and detection. Information on analytical standard used, and providing the chromatograms for standards and samples (e.g. ID confirmation of mixture components), calibration curves, linearity, sensitivity, etc. (e.g. specific further elaboration on the last bullet: “all information relevant for the interpretation of the results, in particular with regard to impurities and physical state of the substance”). Greater/clearer inclusion of analytical data should be encouraged within TGs. For example, a study on a chemical known to be a mixture of two components, with two peaks in the standard chromatogram, however, sample chromatograms cut off after only one peak, highlighting a potential issue with this study that would be missed without submitted chromatograms.



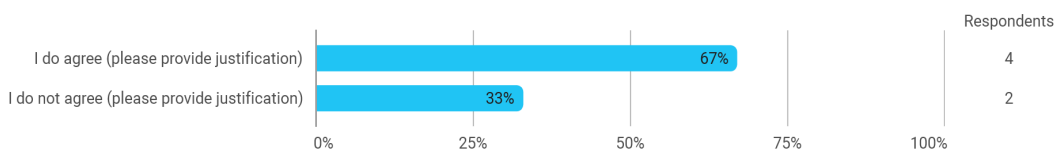
Agreement justifications

- ▶ though I did not recognize this yet a problem because I know only studies where these information are provided
- ▶ same as answer above
- ▶ analytical validity criteria should be given to fulfil requirements of SANTE 10254/2021

Disagreement justifications

- ▶ this is a general remark and mandatory for all studies conducted for regulatory purposes

4. Consider using Test No. 120 for polymers



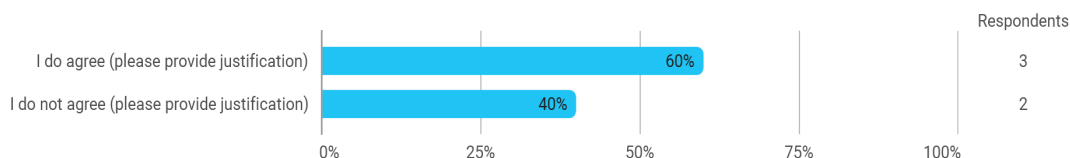
Agreement justifications

- nice link to be placed in para 1
- This may become a regulatory requirement for polymers

Disagreement justifications

- this is a separate guideline and task and should not be mixed with the determination of a water solubility of a substance

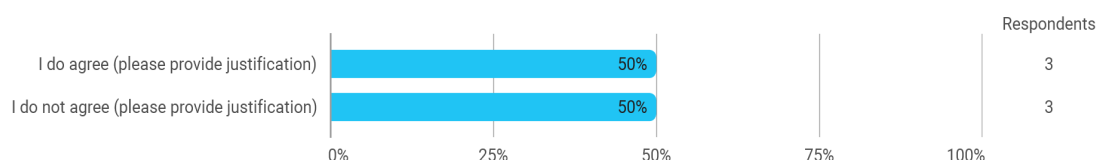
5. Define sample preparation



Agreement justifications

- should be given as essential information when providing study reports though I think that this information should be part of good reporting practice
- additional guidance for difficult substances are necessary

6. Define LOD



Agreement justifications

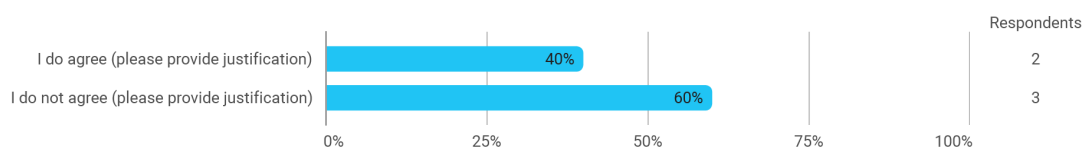
- LOD should be given to fulfil requirements of SANTE 10254/2021

Disagreement justifications

- LOD is define elsewhere. Some basic knowledge is necessary when conducting these studies.

7. Revise wording - The guideline states "This guideline addresses the determination of the solubility in water of essentially pure substances which are stable in water and not volatile."

- polymers are by definition not pure substances, additionally they often contain impurities
- a clearer definition of "stable" would be needed, e.g. is gelation, film-formation, or swelling, a sign for instability? Is a differentiation between physical effect and chemical reaction (e. g. cross-linking due the reaction with the solvent) needed? How to assess polymers with ultrahigh molecular weights?

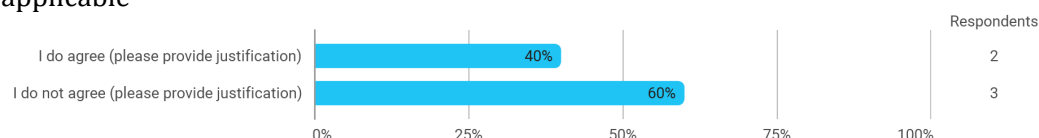


Agreement justifications

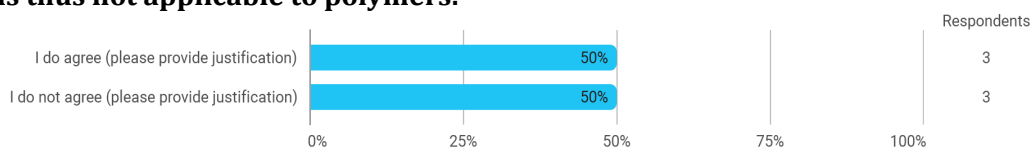
- I think this guideline have its focus first on testing of pure substances. OECD120 more appropriate

8. Revise test method:

- a) The dissolution speed of the polymer may be very low, despite the high polymer solubility (solubility rate for completely soluble polymer is driven be the molecular weight). Which renders the Preliminary test impractical.
- b) For dispersions / emulsions or polymers as very small particles: How to differentiate between dilution and dissolution?
- c) If a water soluble polymer contains impurities which are insoluble, the preliminary test is not applicable

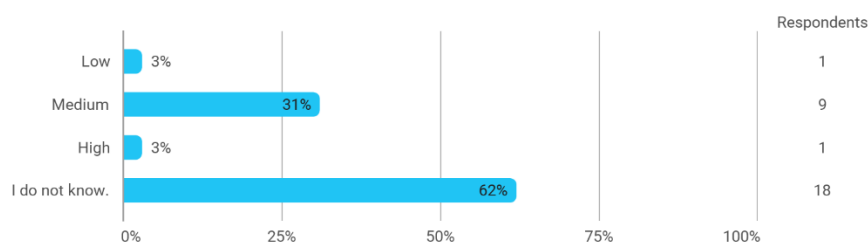


9. **Revise applicability: The TG states: "Under both procedures, the fractions should be checked for the presence of colloidal matter by examination of the Tyndall effect". Even completely soluble polymers lead to the Tyndall effect, this criterion is thus not applicable to polymers.**

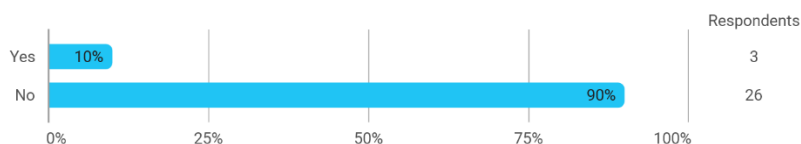


A.2 Test No. 112: Dissociation Constants in Water

Relevance of revision



Interest in being part of an OECD Expert Group

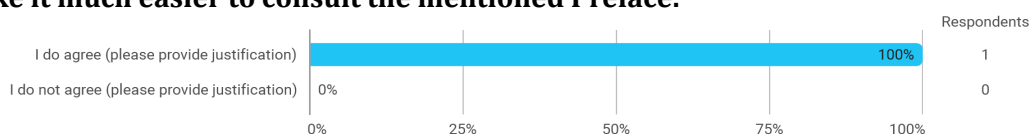


The following suggestions have been collected:

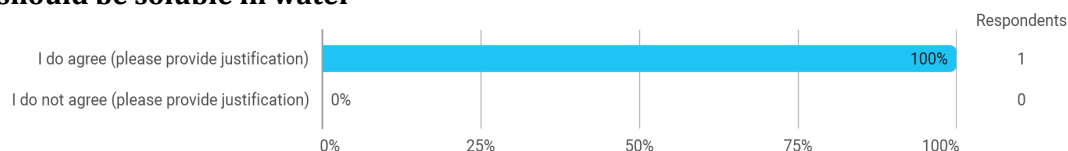
1. Footnote on page 1 - The footnote "Users of this Test Guidance should consult the Preface, in particular paragraphs 3, 4, 7 and 8." mentions the Preface. It is suggested to add a link to the preface document.

Rational: On the OECD website we could not find the Preface document. A link would make it much easier to consult the mentioned Preface.

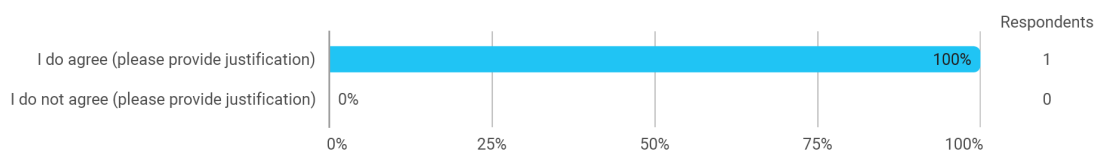
Rational: On the OECD website we could not find the Preface document. A link would make it much easier to consult the mentioned Preface.



2. Use of organic solvent (e.g. DMF, DMSO, NMP) to dissolve the test item and dilution with water e.g. 1:1 (w/w). Potential wording: "The organic solution of the test item should be soluble in water"

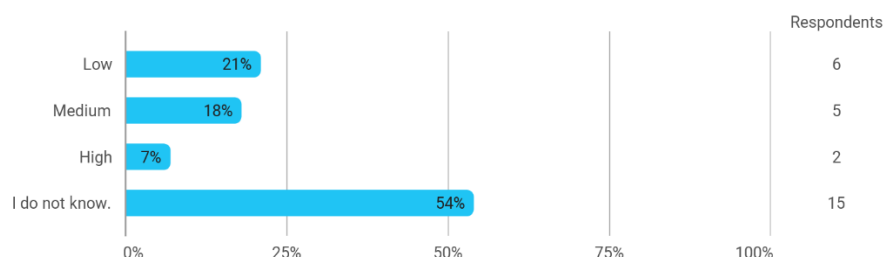


3. Use of organic solvent (e.g. DMF, DMSO, NMP) to dissolve the test item and dilution with water e.g. 1:1 (w/w). Potential wording for (2A): "The dissociation of a chemical in water or a mixture of water and solvent..."
Potential wording for (2B): "...the chemical substance should be dissolved in distilled water or the organic solution of the chemical substance should be dissolved in water." The concentration of the test substance should not exceed the lesser of 0.1 M or half the saturation concentration"

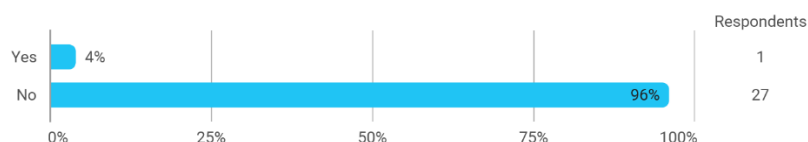


A.3 Test No. 318: Dispersion Stability of Nanomaterials in Simulated Environmental Media

Relevance of revision



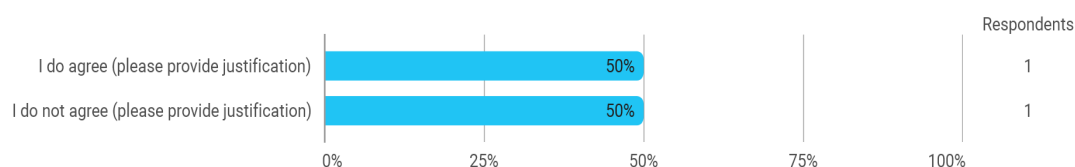
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

4. para 2 - The term "dissolved natural organic matter (NOM/DOM)" may be explained more in detail.

Rational: It seems that natural organic matter (NOM) and dissolved organic matter (DOM) appear in different sections in the document. It may be better to define NOM and DOM more clearly.

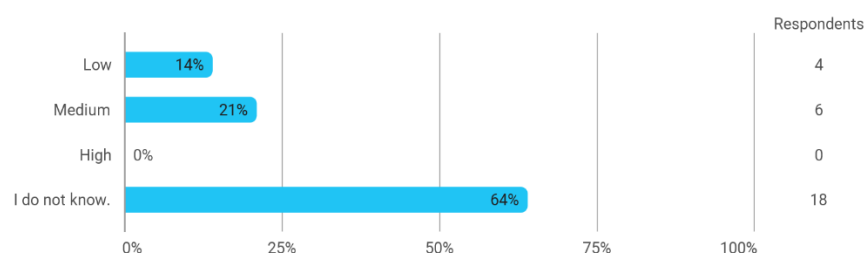


Disagreement justifications

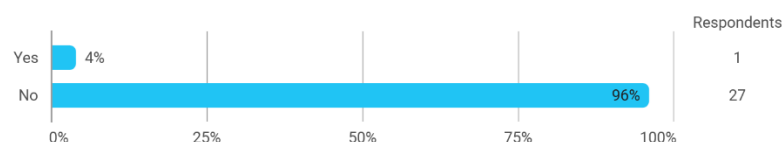
- further information can be drawn from other literature

A.4 Test No. 113: Screening Test for Thermal Stability and Stability in Air

Relevance of revision



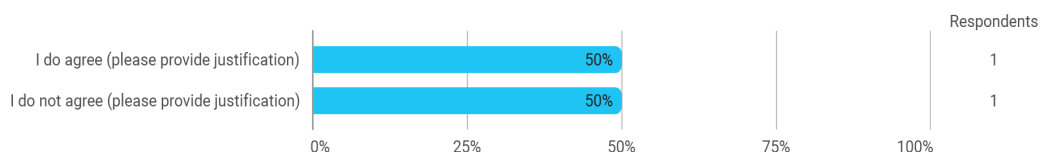
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. para 1- As effects of impurities need to be considered, it seems difficult to test multi-constituent substances or UVCB substances. Therefore, a sentence should be added saying that: "this test is only applicable for defined substances or mixtures of a known composition."

Rational: If substances having several or even unknown constituents it seems to be impossible to assign an observed thermal effects (in form of a peak) to the correct constituent. Next to this, the result of the test cannot be interpreted for the substance as such, but only for specific constituents of it. This should be somehow mentioned in the text of the guideline.

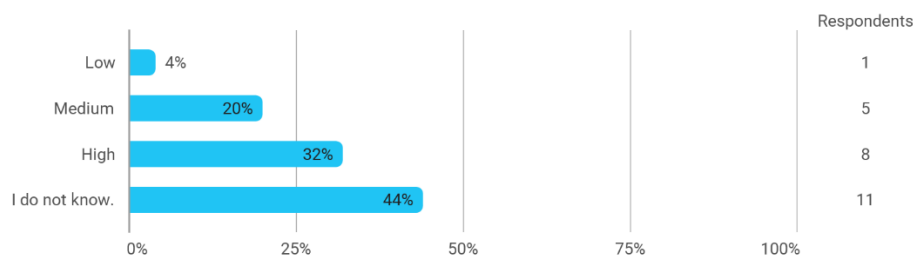


Disagreement justifications

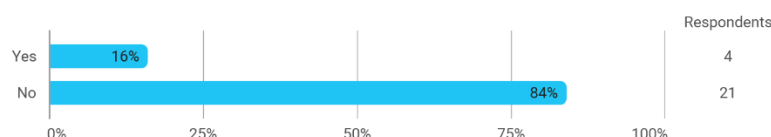
- In my experience the test CAN be usefully used for MOCS; hence the suggestion is incorrect.

A.5 Test No. 107: Partition Coefficient (n-octanol/water): Shake Flask Method

Relevance of revision

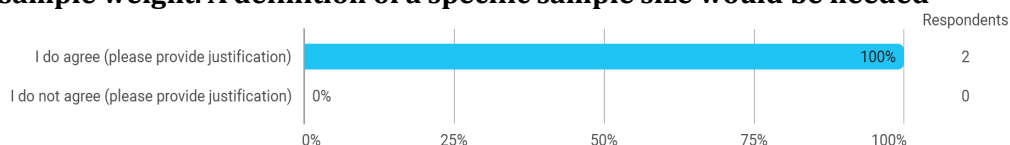


Interest in being part of an OECD Expert Group



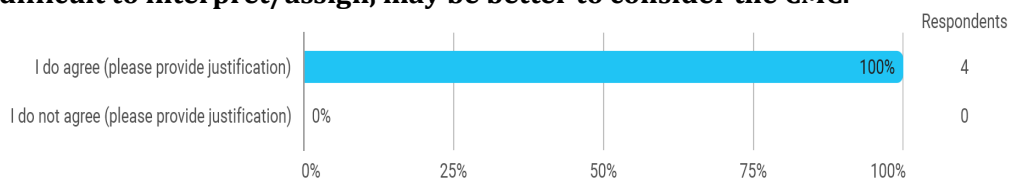
The following suggestions have been collected:

- Adapt preparation instructions: Measured polymer concentration in the respective phase (and hence also the log Kow) can be dependent on the original sample weight. A definition of a specific sample size would be needed**



Agreement justifications

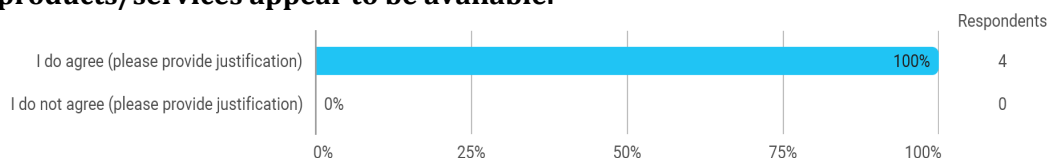
- ▶ general consideration for polymers could be added
- “Pow can be calculated from individual solubilities in water and n-octanol” may not be ideal advice for surface active substances. Perceived water solubility can increase "infinitely" above the Krafft point, thus an "individual solubility" can be difficult to interpret/assign, may be better to consider the CMC.**



Agreement justifications

- ▶ CMC is the nearest proxy for water solubility of surfactants. It is dependent on ionic strength and temperature and this must be taken into consideration when conducting measurements.
 - ▶ I agree
 - ▶ CMC procedure already in use, thus guidance should be adapted
- On that note, a TG for determining CMC could be developed to move forward the state of the science for surface-active substances. Currently the concept only seems to be introduced in STA 23 (from this vantage point...). CMC is increasingly important and seems to be increasingly claimed as an Annex XI Section 2 water**

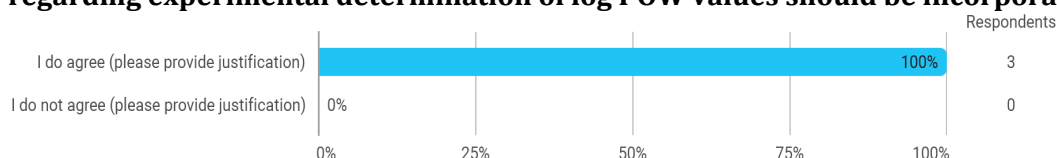
solubility alternative and increasingly suggested as an R7. a endpoint alternative for water solubility. A TG and some guidance on the most common tensiometric, electrophoretic, spectroscopic, etc. methods could be valuable in the identification and evaluation of surface-actives. A number of reviews and commercial products/services appear to be available.



Agreement justifications

- Where the Kraft point allows a CMC measurement to be made this should be encouraged for surfactants. For very high chain length substances this may not be the case.
- I agree
- CMC procedure already in use, thus guidance should adapted

4. para 11- The choice of volume ratio and substance quantity should be harmonized with more recent methods e.g. OECD 123 (Slow-Stirring Method), i. e. LOD/LOQ values should be taken into account. Additionally, it would be helpful to provide tables similar to Annex I of OECD 123 that help estimation of the volumes needed. Rational: The description of the test conditions is very general and depending on many factors. Also insights from the development of the more recent method OECD 123 regarding experimental determination of log POW values should be incorporated.

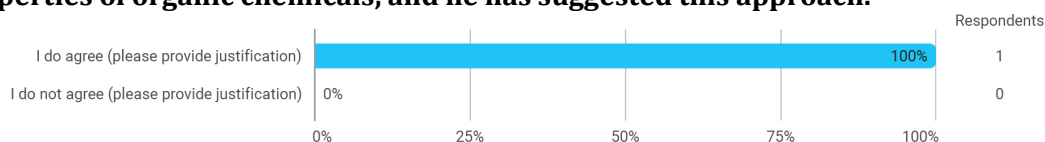


Agreement justifications

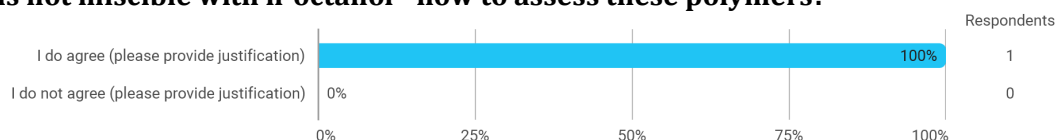
- This sounds a sensible suggestion
- guideline should be more precise in this point

5. general: There are “new” approaches to correct physical-chemical data for water solubility, vapor pressure, solubility in octanol, and partition coefficients for systems of air/water, octanol/water, and octanol/air by using “intra-thermodynamical” relationships between the properties (Beyer et al. 2002; Shen & Wania 2005). These guideline studies overlapping properties, in essence the activity (coefficients) of a chemical in water, air, and octanol. Researchers suggest that when there is more confidence in one result (for instance vapor pressure and octanol solubility from separate tests), then a direct measurement could be conducted or corrected of the octanol–water partition coefficient using the data with the largest confidence in.

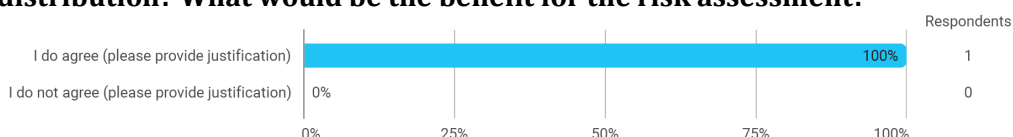
Rational: Professor M. has spent much of his career in determining physical-chemical properties of organic chemicals, and he has suggested this approach.



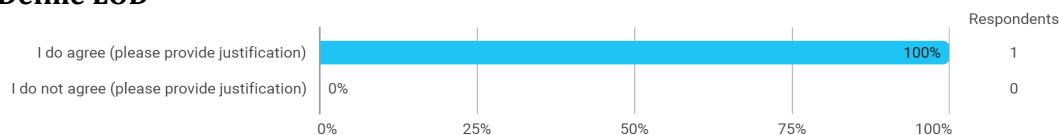
- 6. Revise applicability: Polymers often occur in mixtures where the polymer cannot be isolated. It might be difficult to distinguish between effects caused by the impurities or by the polymer. For high molecular weight polymers, the equilibration time might be very long. In practice, the partition coefficient for polymers is calculated from the individual solubilities in water and n-octanol. However, this is problematic for aqueous dispersions where the aqueous solvent is not miscible with n-octanol - how to assess these polymers?**



- 7. Potential addition to the test method: Solubility might be dependent on chain length, should logPow consider this, e.g. Perform GPC of both phases after partition and compare the molecular weight distribution in each phase? Could a log Kow be calculated for certain peak slices to account for different homologue distribution? What would be the benefit for the risk assessment?**

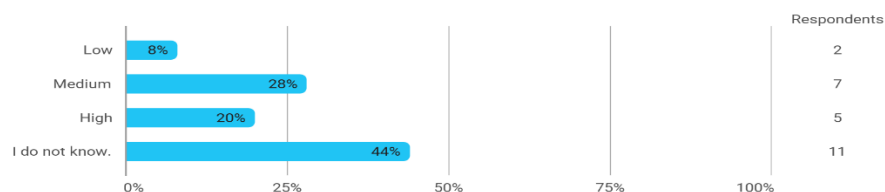


8. Define LOD

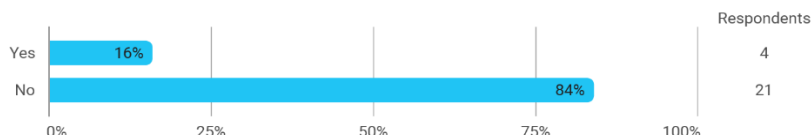


A.6 Test No. 117: Partition Coefficient (n-octanol/water), HPLC Method

Relevance of revision



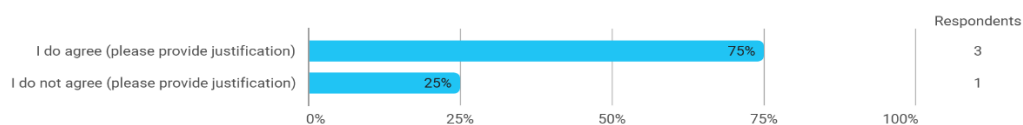
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

- para 11 - A statement should be added to what extend this method is applicable to multi-constituent substances of known or unknown/variable composition.

Rational: Multi-constituent substances and UVCB substances are more common in chemistry regulation nowadays. Testing methods should be revised regarding their compatibility with such substances.



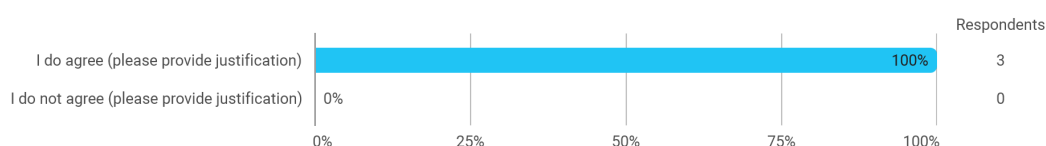
Agreement justifications

- ▶ The guideline is applied to UVCBs so it is a good idea to make this clear for all users.
- ▶ although it will be hard to give guidance to the broad range of multi-constituent substances

Disagreement justifications

- ▶ The guideline contains all elements to use this method also for UVCBs under the prerequisite that the substances are in the scope of the application

2. Revise applicability and define standards for polymer assessment

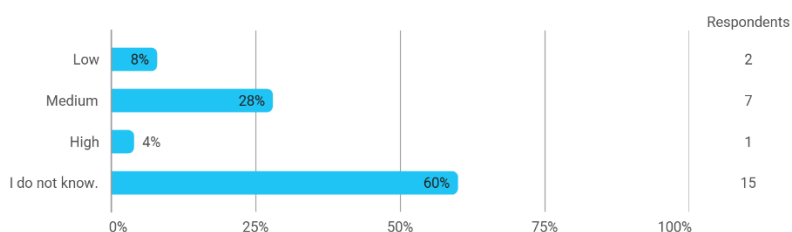


Agreement justifications

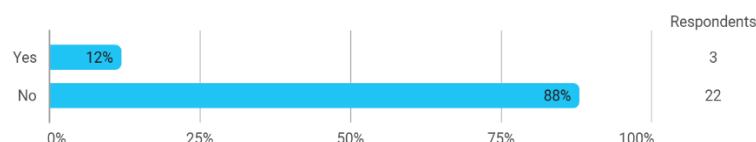
- ▶ I agree in principle. It should be generally clarified which approaches are useful to determine the log Kow for polymer fractions.
- ▶ General update of OECD methods for polymers could be useful

A.7 Test No. 123: Partition Coefficient (1-Octanol/Water): Slow-Stirring Method

Relevance of revision



Interest in being part of an OECD Expert Group



The following suggestions have been collected:

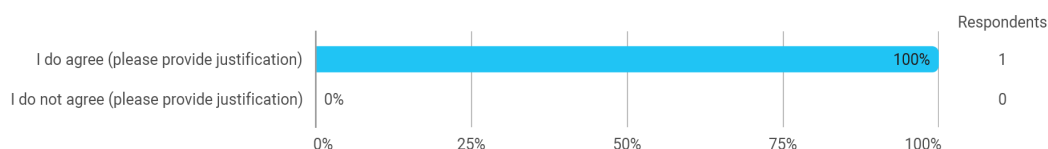
1. para 2 - For the determination of Kow, only the slow-stir and the HPLC method are discussed. There is an alternative, direct determination that avoids the problem of microemulsions and is considerably faster than the slow-stir (also direct) method and gives values in accordance with the direct method, in disagreement with the HPLC method (indirect). It uses a semipermeable membrane to keep the two phases

separated during equilibration. Two publications describe the method:

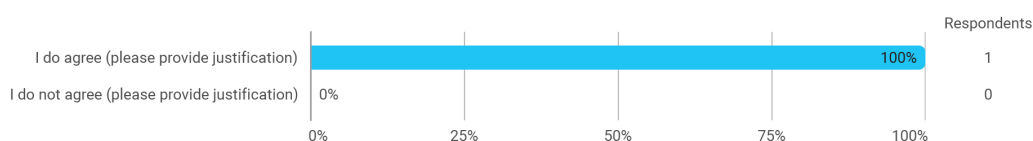
Anal. Chem. 71 (1999) 3610-3614

J. Pharm. Sci., 90 (2001) 1948-1954

Rational: The proposed method is vastly faster than the other direct method (slow-stir) discussed in the OECD guideline, working in hours instead of days. The values thus obtained correlate in an excellent fashion with values obtained through the slow-stir method.

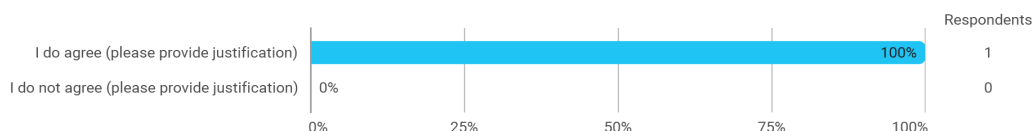


2. para 27 - According to paragraph 27, the minimum number of analyses required is indicated to be 5 (n=5). However, paragraphs 39, 47 and 50 all mention 4 samples. Please correct paragraph 27.

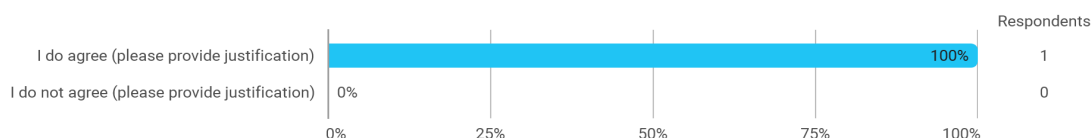


3. ANNEX 1 Table: Computation of volumes: The left column of the given table is titled "log Kow"; this should be changed to "log Pow".

Rational: Since the rest of the guideline only uses "log Pow", it should be used in this table as well for consistency reasons. Therefore, "log Kow" should be replaced using "log Pow" instead.

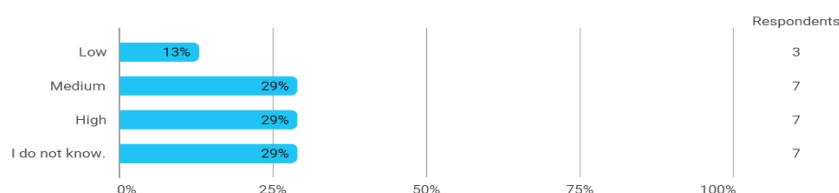


4. Define LOD: 'Limit of detection for polymer analysis may be much higher compared to defined substances because they appear in distributions and not as a single defined molecule.

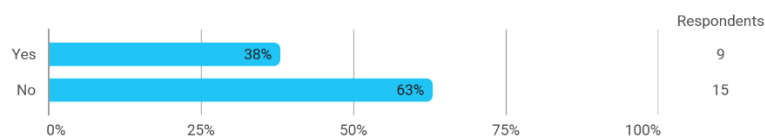


A.8 Test No. 106: Adsorption -- Desorption Using a Batch Equilibrium Method

Relevance of revision

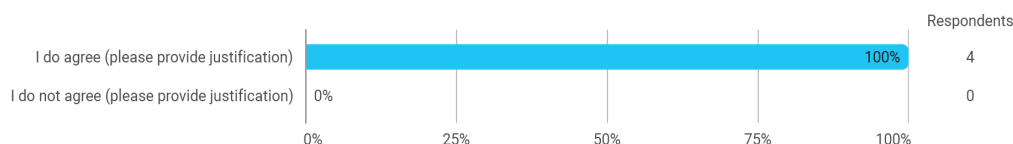


Interest in being part of an OECD Expert Group



The following suggestions have been collected:

- General: The Guideline uses "cm³" all through the document. System International d'unites (SI) recommend "ml".**

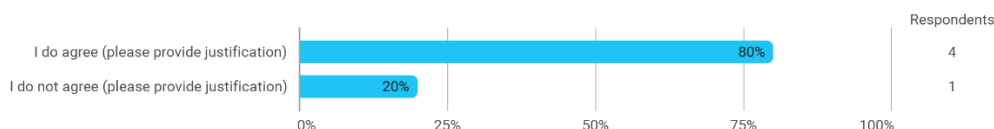


Agreement justifications

- no further comment
- Agreed
- SI units should be preferred

- para 3 - Investigation on desorption kinetics/Freundlich isotherms should be optional.**

Rational: The desorption endpoints are not used in modelling calculations for risk assessments.



Agreement justifications

- Optional agree for Desorption and Isotherms. Furthermore, validity of desorption using the direct method should be addressed as well as the performance of the test
- Desorption parameters are not used in leaching models, therefore there is no need for them to be mandatory
- The need for desorption kinetics may depend on the regulatory framework in which the study will be used so should be optional. In addition desorption kinetics may be unfeasible in certain circumstances (for example very short lived or highly sorbing substances)

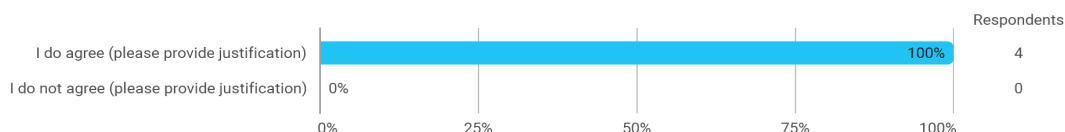
Disagreement justifications

- Do not think this is useful. It is the aim of the guideline to give a comprehensive overview about what is possible. In my opinion, the relevance for the conduction of the desorption experiments is triggered by the regulatory framework requesting this study (e.g. pesticides vs. industrial chemicals)

- para 5 - Conduction of the study with other matrices, e. g. activated sludge**

Rational: According to the new guidance document for ERA of human pharmaceuticals by EMA OECD106 is the preferred method to study the adsorption/desorption

behaviour of a substance. Investigation of adsorption/desorption on two types of activated sludge are required according to the EMA guidance. But, conduction of the OECD106 with activated sludge is not described in the corresponding guideline. This should be included and a harmonised procedure should be proposed.

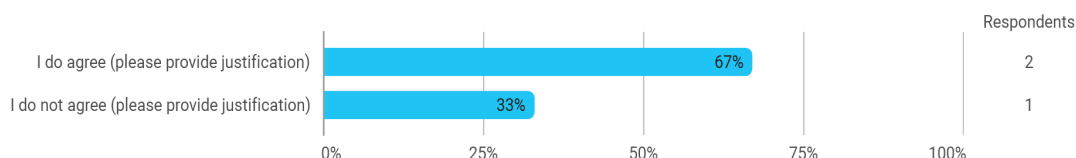


Agreement justifications

- preparation of sludge as well (e.g. OPPTS)
- Agreed
- relevant e. g. for pharmaceuticals according to EMEA/CHMP/SWP/4447/00 Rev. 1

4. para 39 - Footnote 1

The equation written in the footnote should be lifted into the main text. This because the very definition of Kd is what is told here. This is perhaps the most important information for a reader to understand the scope of the guideline.



Agreement justifications

- no further comment

Disagreement justifications

- is okay in the existing guideline version

5. para 39 - More information needed

The main text could also be complemented with information describing the algebraic relation between Kf and Kd. For instance like this:

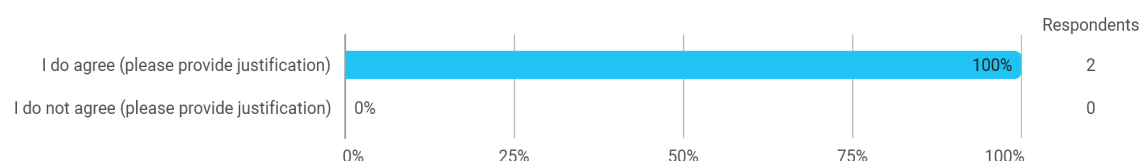
The Freundlich constant is per definition: $KF = CS / (CW)^n$

This means that: $CS = KF \times (CW)^n$

The distribution coefficient is per definition: $Kd = CS / CW$

Replacing CS by $KF \times (CW)^n$ gives: $Kd = KF \times (CW)^n / CW \Rightarrow Kd = KF \times (CW)^{(n-1)}$

which is an equation that is not explicitly given in the current version of OECD 106, but can be found as Equation 9-8 in Schwarzenbach et al., 2003.

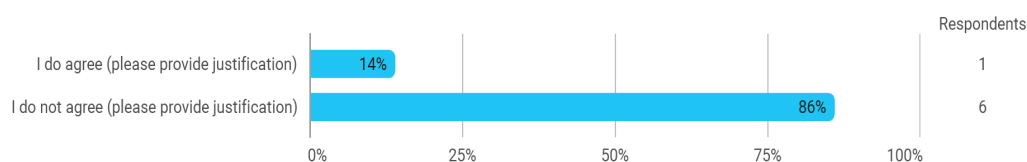


Agreement justifications

- in combination with proposal in 4 useful

6. paras 43, 44, Annex 5. Colloids

These paragraphs and the appendix indicate colloids play an important role in the sorption process. It is trusted that this is also a very widely accepted hypothesis among researchers (e.g. Burkhardt, 2000; Govers & Krop., 2001). Therefore, it would be appreciated if a revised version of the OECD 106 clearly includes analytical determination of the amount colloids in the filtered water, or in the supernatant (when centrifugation is used). The proposal is that standard elemental analysis of total organic carbon is included as an auxiliary parameter to report. Researchers and authorities will thereby be given better data for improving their understanding of DOC. This is extra important now, since more and more laboratories are starting to use ultracentrifugation, where colloids are removed, and thereby generates OECD Koc values with significantly lower DOC content (possibly giving higher Koc values).



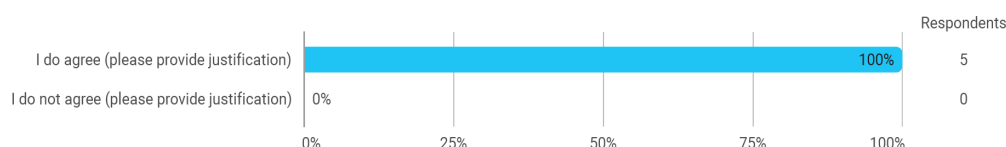
Disagreement justifications

- ▶ Determination of colloid concentration would be specialized analysis and not possible for many laboratories. Furthermore, the concentration of dissolved colloids would be very low in comparison to the amount of bulk soil, so their influence would likely be negligible. Ultracentrifugation would only be used for compounds with extremely low sorption, where the influence of colloids would be even less
- ▶ perhaps better to define the equipment which should be used to maximise comparability
- ▶ The colloid fraction is such a low percentage of the test system that measurements of adsorption to this fraction are not technically feasible. Such measurements is not necessary because the amount of colloid compared to the soil left after centrifugation will be negligible and thus within the range of overall experimental accuracy. Ultracentrifugation is expected to be used only for very low adsorbing compounds to separate phases as best as possible. Such compounds will have a colloidal bound fraction that will be negligible to the overall adsorption coefficients.
- ▶ Determination of colloid amounts and the adsorbed parts of test items adsorbed from low liquid volumes is not possible. And even more, it is not necessary because the amount of it compared to the soil left after centrifugation will be negligible and thus within the range of overall experimental accuracy. Ultracentrifugation is expected to be used only for very low adsorbing compounds to separate phases as best as possible. Such compounds will contribute if colloidal bound even less to the overall adsorption coefficients.
- ▶ Disagree. The exact determination of colloid amounts and even more, the adsorbed parts of test items adsorbed thereon from low liquid volumes is not possible. And even more, it is not necessary because the amount of it compared to the soil left after centrifugation will be negligible and thus within the range of overall experimental accuracy. Ultracentrifugation is expected to be used only for very low adsorbing compounds to separate phases as best as possible. Such compounds will contribute if colloidal bound even less to the overall adsorption coefficients.

- Exact determination of colloid fractions and, even more, parts of test items adsorbed onto colloids when applying low volumes of the liquid phase is not feasible. Moreover, it is not needed since the respective amount of it compared to the soil left after phase separation will be negligibly small and thus within the range of the overall experimental accuracy. Ultracentrifugation is expected to be used only for very low adsorbing test items to separate soil and liquid phase as best as possible. These test items will contribute, if colloidal bound, even less to the overall adsorption coefficients. Therefore, The exact determination of colloid amounts and even more, the adsorbed parts of test items adsorbed thereon from low liquid volumes is not possible. And even more, it is not necessary because the amount of it compared to the soil left after centrifugation will be negligible and thus within the range of overall experimental accuracy. Ultracentrifugation is expected to be used only for very low adsorbing compounds to separate phases as best as possible. Such compounds will contribute if colloidal bound even less to the overall adsorption coefficients. Therefore, elemental analysis of total organic carbon should be optional depending on test substance and centrifugation and round force used for separating soil and supernatant (e.g. ultracentrifugation).

7. **para 64 - The term “mass balance” should be clarified to avoid confusion. OECD106 describes this as the recovery of the test item, but mass balance usually refers to recovery of total 14C. “Test item accountability” would be a clearer term or at least definitions of parental mass balance and mass balance (for labelled test items recommended) should be added**

Rational: To avoid confusion with recovery of total radioactivity which is referred to as mass balance in other Environmental Fate Studies

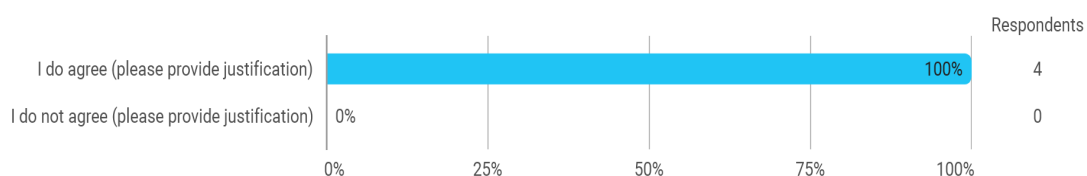


Agreement justifications

- no further comment
- This would reduce confusion when assessing study quality using the EFSA reviewers checklist
- clarification useful
- Agree, some additional clarity could be provided around this point.

8. **para 64 - Add the option to perform the adsorption test with sterile soils to increase the stability of the test items, and include guidance which sterilisation methods should be avoided (e.g. because they may impact sorption-relevant soil structures and constituents).**

Rational: If the instability of the test item is due to biological degradation, running the test with sterilized soil can be an option to overcome the stability issues. However, the sterilization methodology should be carefully selected to avoid relevant changes of the soil. We recommend sterilization with gamma irradiation.

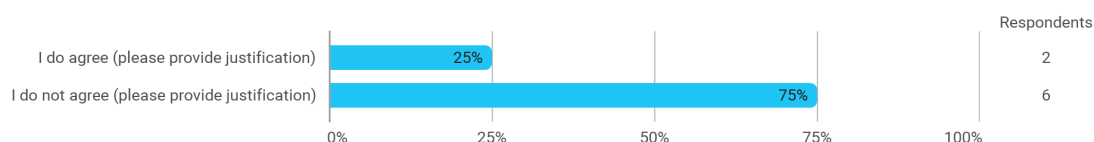


Agreement justifications

- Sometimes this is necessary anyway to improve stability, therefore it is a valid option provided that non-destructive methods of sterilization are used (e.g. gamma irradiation)
- relevant remark
- Agree, clear guidance as to appropriate methods or sterilisation (or which to avoid) would be helpful here.

9. para 86 - Koc or better Doc

It might be relevant to change to the notation log Doc (log organic carbon – water distribution ratio). This is used in modern textbooks on environmental organic chemistry (e.g. Schwarzenbach et al., 2003, page 321) in order to underline that it is a distribution ratio of (possibly) many sorbed and dissolved species. The same argumentation for why Dow should replace Kow (octanol – water) is used. The current “Koc” values are clearly Doc in this sense, since they change with pH (indication the sorbing molecule species change composition).



Agreement justifications

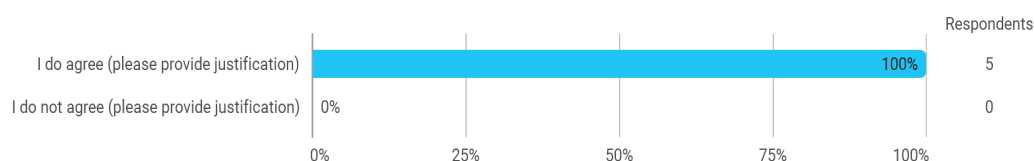
- conditions of the appropriate sterilisation processes should be stated as well

Disagreement justifications

- DOC is a term currently used for dissolved organic carbon. Therefore replacing Koc with Doc would simply add confusion
- Koc okay, not a necessary change
- stability preliminary tests are run to ensure that there are not multiple sorbing species. It is also not valid to state “The current “Koc” values are clearly Doc in this sense, since they change with pH (indication the sorbing molecule species change composition).” – many molecules do not exhibit pH dependent adsorption.
- K(oc) is a well introduced term associated with adsorption studies and Freundlich partition theory. Doc could be misleading while unintentionally written in capital letters (DOC- Dissolved Organic Carbon), it means something different
- Koc is a well established term associated with the determination of adsorption behaviour, incl. Freundlich partition theory. Doc could be misleading while unintentionally written in capital letters (DOC-Dissolved Organic Carbon), it means something different. Although the suggestion might be understandable from a scientific point of view, the benefit of changing the established naming is highly questionable.

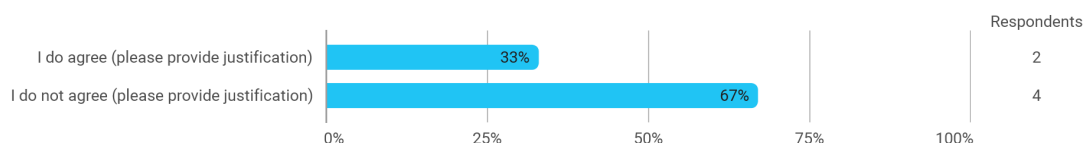
- Disagree, Koc is well understood in this context. In addition Doc could create confusion with Dissolved Organic Carbon.

10. para 87 - (Page 16, line 7). "Freudlich adsorption coefficient" is misspelled and should read "Freundlich adsorption coefficient".



11. paras 87 and 94 (Page 16, line 8 & page 18 line 20).

The units of Kf could, for reasons of clarity, be expressed in " $(\text{mg/kg})/(\text{mg/l})^{1/n}$ " or " $(\mu\text{g/kg})/(\mu\text{g/l})^{1/n}$ " instead of " $(\mu\text{g l}^{-1/n} \text{ cm}^3)^{1/n} \text{ g}^{-1}$ " which is a shortened unit and somewhat cryptic. In many DARs/EFSA-reports and other EC-literature, the units of Kf are frequently left out, or " l/kg " is used, which is incorrect since this is only valid when $1/n = 1$. Perhaps this can be avoided if using more comprehensible units. Within the modern environmental research literature (e.g. Schwarzenbach et al., 2003), the notation " $1/n$ " is abandoned, and instead only " n " is used. These are synonymous. The notation $1/n$ may cause confusion since it may be interpreted as an inverse, which it is not. It seems wise to change-over to n .



Agreement justifications

- could be more straightforward

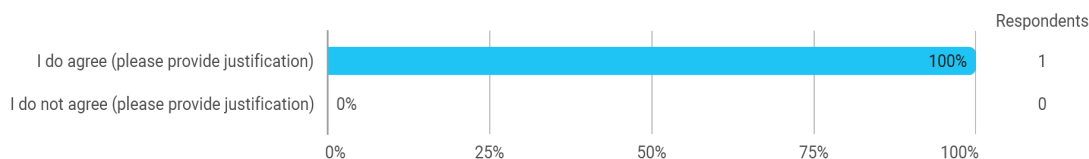
Disagreement justifications

- Although this may be technically correct, it would just add confusion and disrupt the flow of the text in study reports. Therefore the simpler units should be retained
- The current simplified units are globally well recognized and used for risk assessments of PPPs, therefore a change is not warranted and would be highly disruptive
- Although it might be scientifically reasonable the benefit of changing the established units and naming is highly questionable. With regard to the unit of the Kf value it is recommended to keep the simplified unit of L/kg as Kf always has to be considered with its respective $1/n$ value.
- The units in this context are well understood and Kf is always paired with the respective $1/n$ value.

12. para 95 - Test report

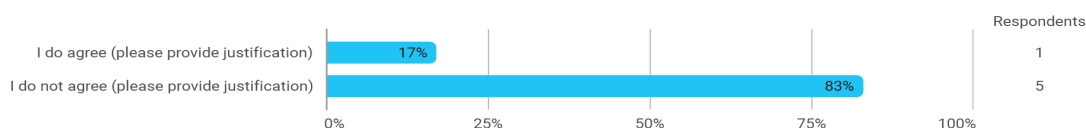
It could be added an explicit request that a reference value for the dissolved solute concentration (C_e) must be stated (reported) when the distribution coefficient K_d is calculated from non-linear sorption isotherms. For a non-linear sorption isotherm, Kf is constant (with its specific units) for all concentrations, but K_d is not. This is important since, for example, if two K_d values (A & B based on Kf values with different

1/n) are compared at 1 mg/l, the value A may be higher than B, while the opposite may be true at the (commonly much lower) environmental exposure concentration of interest in the risk assessment. The examined range of concentration should also be a specific request to report. This because the sorption isotherm may follow different fits (Freundlich, linear, etc) in different concentration regimes.



13. para 95 - Test report

There are some new ideas on how to derive more “unit equivalent” Kf values (‘Kf) by expressing the sorption isotherm with the equation $CS = 'KfCr_n$, where Cr (“reduced concentration”) is the equilibrium solution concentration normalised to the (subcooled) liquid aqueous solubility (Carmo et al., 2000). There is interest in this approach because it could give more “inter-compound comparable” data. Although the OECD guideline 106 never have recommended comparing traditional Kf values with different 1/n, this is frequently done within risk assessment, but is clearly incorrect since they have different units (like comparing m2 with m3) and sometimes misleading. Have the ideas by Carmo et al., or similar approaches, been discussed by the Guideline work group?



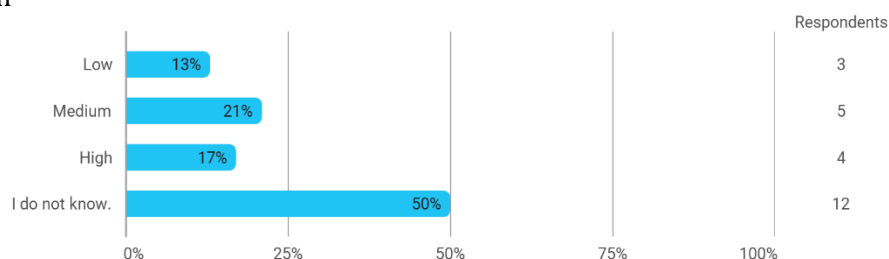
Disagreement justifications

- ▶ Comparison of Kf across compounds is not a key output of the study. If anything, comparison on groundwater predicted concentrations is more valuable, which incorporates the contribution of both Kf and DT50, through specific regulatory models for different global regions. As such this is not necessary. Furthermore, there are concerns regarding the reliability of the method reported by Carmo et al., such as the use of very low soil masses (50-200 mg) which would introduce more error into the measurement
- ▶ The approach outlined by Carmo et al is not applicable to PPP's. Relying on a unit conversion method that was derived from procedures not in line with the OECD 106 guideline is flawed.
- ▶ this approach by Carmo et al. and derived endpoints are not compatible with the current risk assessment procedures for PPPs. Applicability for compounds other than HOC's was not investigated and the used number of soils appeared to be not sufficient to derive universal validity. Furthermore, described experiments are lacking severely of OECD 106 guideline compliance. E.g. use of saturated stock solutions, use of an additional, hazardous compound, questionable representative soil aliquots of 50-200 mg, no examination of the soil phase after phase separation and absence of fortified control samples, validating analytical determination procedure from fortification to instrumental measurement with limited UV detection selectivity etc.

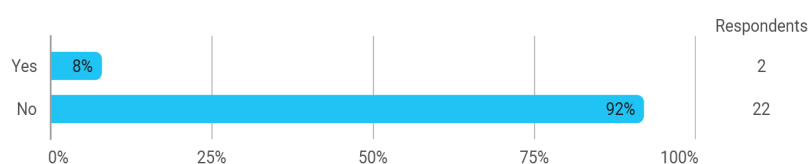
- Disagree; this approach by Carmo et al. and derived endpoints are not compatible with the current risk assessment procedures for PPPs. Applicability for compounds other than HOC's was not investigated and the used number of soils appeared to be not sufficient to derive universal validity. Furthermore, described experiments are lacking severely of OECD 106 guideline compliance. E.g. use of saturated stock solutions, use of an additional, hazardous compound, questionable representative soil aliquots of 50-200 mg, no examination of the soil phase after phase separation and absence of fortified control samples, validating analytical determination procedure from fortification to instrumental measurement with limited UV detection selectivity etc.
- We disagree. The approach by Carmo et al. and the derived endpoints are not compatible with procedures of the PPP risk assessment procedures. Applicability for a broader range of compounds other than HOC's was not assessed and the used number of soils appeared to be not sufficient to draw a universal conclusion. In addition, the experiments are lacking severely of OECD 106 guideline compliance (e.g. soil aliquots of 50-200 mg, no examination of the soil phase after phase separation and absence of fortified control samples, etc.).

A.9 Test No. 121: Estimation of the Adsorption Coefficient (Koc) on Soil and on Sewage Sludge using High Performance Liquid Chromatography (HPLC)

Relevance of revision

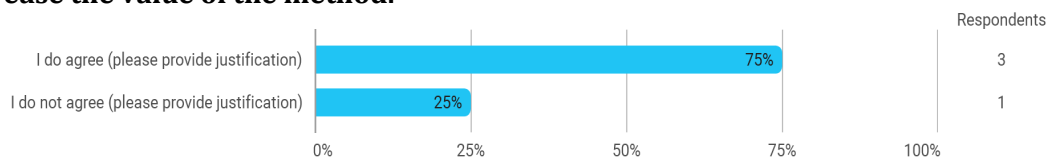


Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. **para 3 - Consideration of additional stationary phase options would be beneficial for those compounds for which the cyanopropyl column is not suitable.**
Rational: I am led to believe that the cyano column can provide a reasonable estimate for neutral compounds, but it less good for ionisable compounds. More options could increase the value of the method.



Agreement justifications

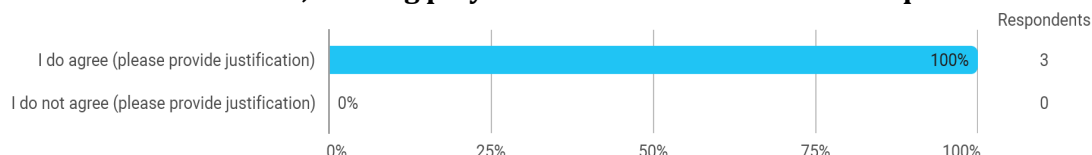
- no further comment

- This would significant broaden the applicability of the guideline, although it is rarely used in the agrochemical industry (as OECD 106 is a required study)

Disagreement justifications

- based on my experiences this column has a broad application range

2. Revise applicability and define standards for polymer assessment: a) Aqueous dispersions: aqueous solvent is not miscible with n-octanol - isolation of polymer sensible / possible? Is the correlation applicable at all for polymers? Many polymer products cannot be assessed chromatographically following the HPLC conditions mandated in OECD TG 117; Missing polymeric reference standards for quantification

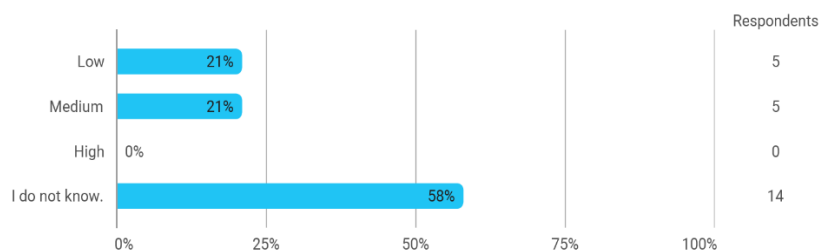


Agreement justifications

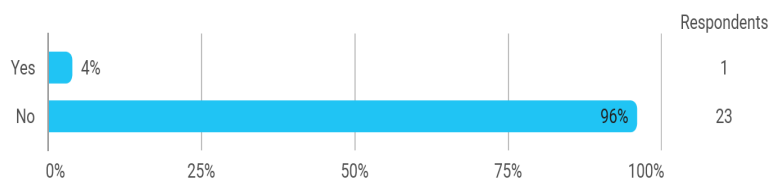
- relevant question how this guideline (Koc) and those for log Kow can be adapted to get meaningful results for polymers
- It is a good idea to give advice on testing polymers, but the suggestions should be double checked to decide if they are valid

A.10 Test No. 312: Leaching in Soil Columns

Relevance of revision

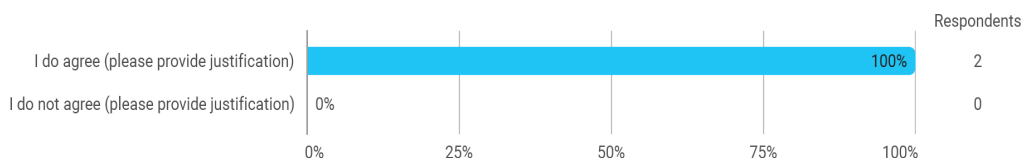


Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. para 46 - Would be useful if more details or equations can be given on converting the experimental leaching distance to equivalent soil KOC.
Rational: To ensure consistency of approach across users

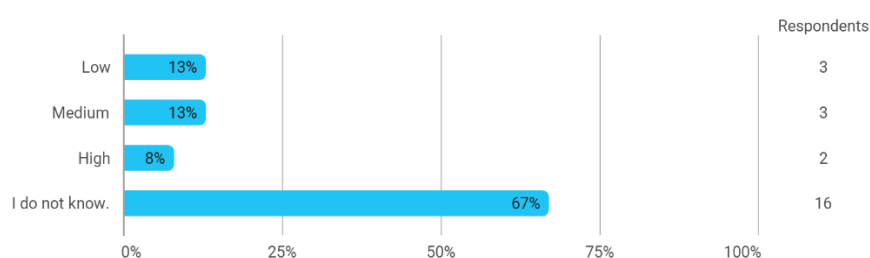


Agreement justifications

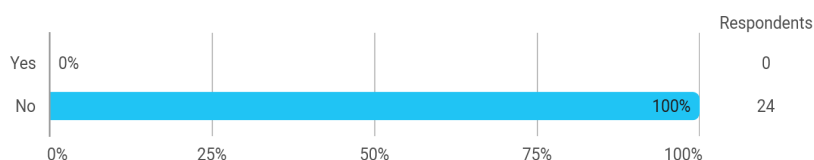
- It has often been stated that this study can be used to estimate Koc when OECD106 is possible, but it is not really defined how this should be done. Therefore, this clarification would be helpful

A.11 Test No. 122: Determination of pH, Acidity and Alkalinity

Relevance of revision

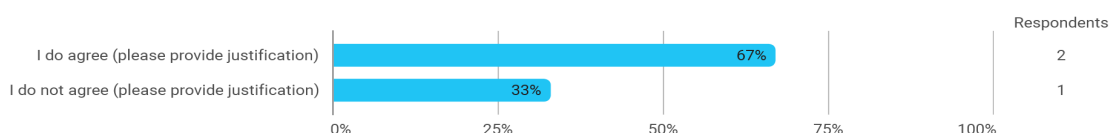


Interest in being part of an OECD Expert Group



The following suggestions have been collected:

- 1. Use of organic solvent (e.g. DMF, DMSO, NMP) to dissolve the test item and dilution with water e.g. 1:1 (w/w) in order to enable use for polymers.**



Agreement justifications

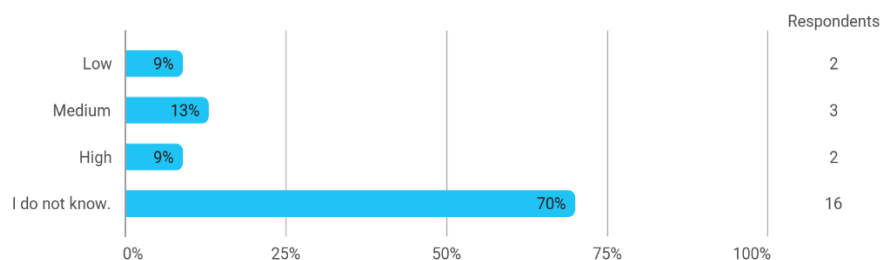
- polymer testing will become increasingly important for REACH registration

Disagreement justifications

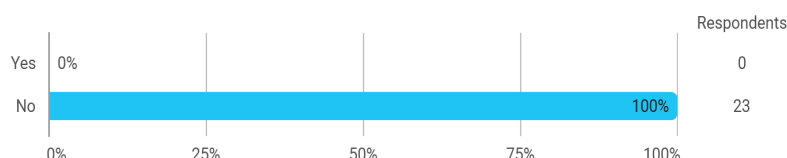
- not necessary

A.12 Test No. 115: Surface Tension of Aqueous Solutions

Relevance of revision

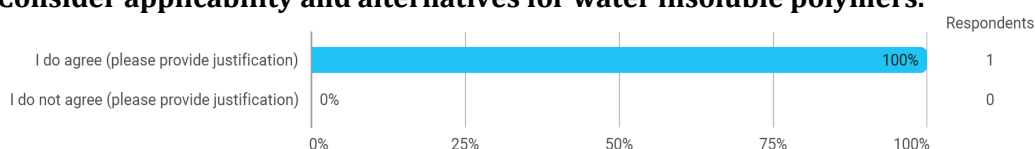


Interest in being part of an OECD Expert Group



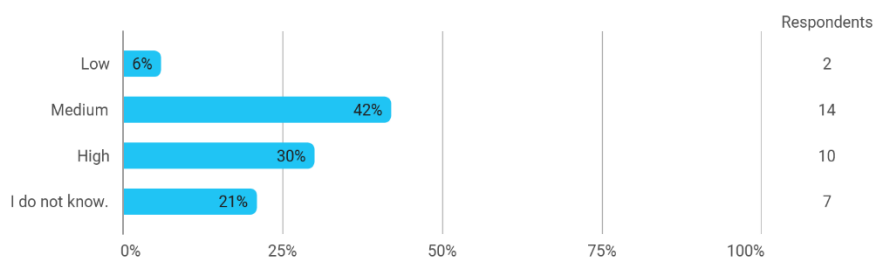
The following suggestions have been collected:

- Revise the suggested methods, consider alternative methods like pendant drop. Consider applicability and alternatives for water insoluble polymers.**

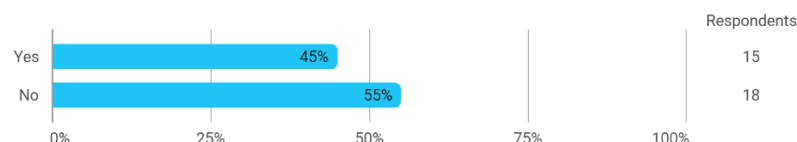


A.13 Revised Introduction to the OECD Guidelines for Testing of Chemicals, Section 3

Relevance of revision



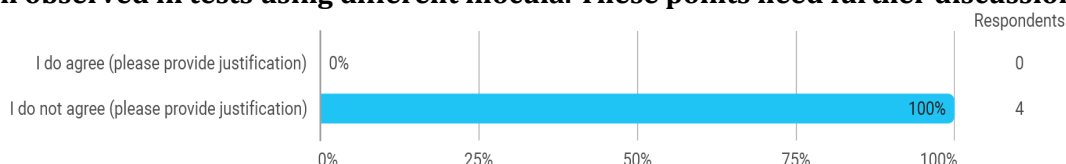
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

- para 6 - It is proposed to remove or amend the italicized sentence. Domestic sewage, activated sludge or secondary effluent is the typical source of microorganisms (inoculum) in tests for ready biodegradability. The inoculum should not have been pre-adapted to degradation of the test substance by previous exposure to the test substance or structurally related chemicals.**

Rational: We would propose that adaptation to the presence of chemicals is a naturally occurring phenomenon that has been observed in almost all environmental compartments and can be induced in the laboratory. However, it cannot be ruled out that microbes in all environmental compartment have been exposed to industrial and naturally occurring chemicals at some time. Therefore, it may be beyond the control of a performing laboratory to verify this. Dalmijn et al. 2020 suggested that including acclimation and adaptation in biodegradability tests may also reduce the variability often observed in tests using different inocula. These points need further discussion.

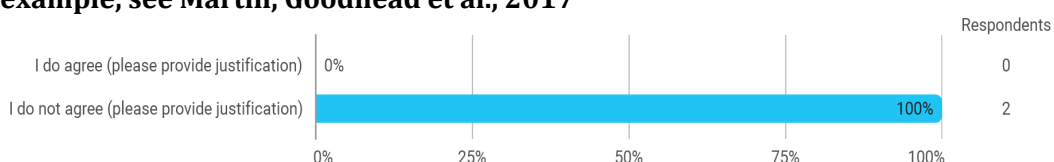


Disagreement justifications

- unclear - no italics?
- Use of non-pre-adapted inoculum is important for regulatory purpose. Therefore, I would not agree with removing this completely. However, more information on the suggestion would be appreciated, for example , how the text should be amended.
- The suggestion does not say what the desired amendment should be, however, if the phrase is changed to allow for adapted sludge, it could result in significantly more materials being classified as readily biodegradable. There needs to be a better understanding of how and why the proposers want this section modified.

2. para 13 - It is suggested that with the OECD 301/302/306 &310 TGs in their current form the following sentence does not generate comparative data that are used to derive the same conclusions. Therefore, it is recommended to perform further work (please see the rationale), "Ready biodegradability tests must be designed so that positive results are unequivocal".

Rational: Currently, there are seven different types of test designs, which are widely used, and inocula can be selected from five different environmental sources with different concentrations and microbial communities (OECD 1992a; OECD 2014; Kowalczyk et al. 2015). The inocula can further be prepared in up to four different ways (Goodhead et al., 2014). Consequently, RBT outcomes for a given substance can differ widely (Thouand et al. 1995; Vazquez-Rodriguez et al. 2003; Goodhead et al. 2014). In order to improve comparability between the outcome of comparative tests a standardised control inoculum could be considered within which the test substance and an agree reference bench mark substance could be compared. The inoculum could be composed of a consortium of known cultured organisms or from a known and characterised natural sample that contains a broad range of enzyme activity. Aspect of improved standardisation to explore include high through-put tests that can be performed in well-plate formats. These allow numerous replicates to be prepared beyond the capabilities of the normal laboratory scale studies, reducing variability. For example, see Martin, Goodhead et al., 2017



Disagreement justifications

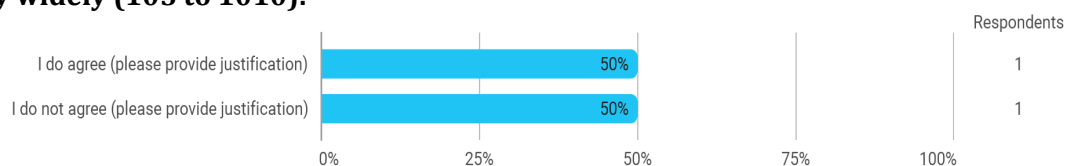
- ▶ I don't fully understand why it should be changed and how. Further information or text suggestion would have been helpful.
- ▶ While I agree that there are some issues with comparing data across the guidelines mentioned, the point of the screening tests is to quickly and cheaply identify materials that need more investigation from those that don't. Thus, it would seem to be important to have unequivocal results. I am interested in seeing the rationale the proposers provided, since it was not included here.

3. **para 20 - This text does not reflect how conservative the inocula density and community size are within these studies. The suggestion is that this is presented in the revision of this introduction and any proposed updates in the test guidelines. " In all the tests, the test substance providing the sole source of organic carbon (except for carbon associated with the biomass) is diluted in a test medium containing a relatively low concentration of biomass. In all the tests, a non-specific analytical method is used to follow the course of biodegradation."**

Rational: It is questionable whether the relatively small sample sizes used for biodegradation tests are sufficient to represent the microbial community encountered by substances freely diffusing throughout a given environment.

Additionally, the variability in the results indicate that this needs to be addressed.

Inoculum density in a ready biodegradability test (RBT) is ~ 10⁴ cells/mL, this is typically ten thousand times less concentrated than the same community in a typical activated sludge wastewater treatment plant (WWTP) (10⁸ cells/mL), and the total number of cells in an RBT test vessel, 108, is a billion times fewer than that of a small sized WWTP (~10¹⁷ in 100 m³ reactor), with ensuing differences in diversity. Across different biodegradation test systems, the cell concentrations (10¹ cells/mL in some RBTs to 10¹⁰ cells/g in soil simulation tests) and total number of microorganisms vary widely (10⁵ to 10¹⁰).



Agreement justifications

- ▶ The intro section and guidelines could use some additional discussion on the conservative nature of the inocula

Disagreement justifications

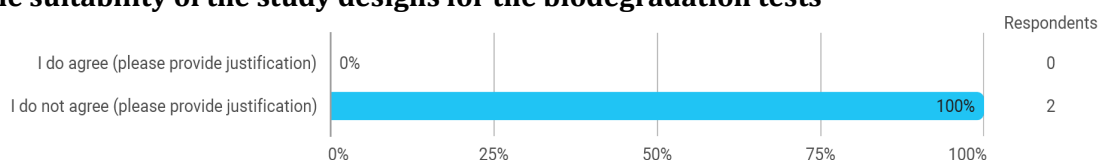
- ▶ para 20 sufficiently describes this topic
 - ▶ Some more reasoning why this change is considered necessary would have been useful.
4. **para 22 - Contradictory statement that influences the test performance and subsequent interpretation of the data. The proposal for this is that the statement below is now out-dated and current cost-effective scientific techniques need to be explored. The comments and rationale that have been included here should be considered to demonstrate that the current methodology and scientific guidance is out of date for any study utilising environmentally derived biomass. The improved**

methodology would enable improved confidence in data that is compulsory for any regulatory conclusion or decision. "It has been recognised that standardisation of the inoculum might also improve the comparability of the methods. However, it was concluded that this is not possible without significantly reducing, at the same time, the number of species present in the test system." This is ~33 years old: "An OECD inter-laboratory 5 comparison exercise (ring test) (1) took place in 1988 in order to ensure the practicability and validity of the tests".

Rational: In light of the review by Kowalczyk et al. 2015, which highlights the impact of inter- and intra-laboratory variation in biodegradation studies and recent academic publications i.e. Ott et al. 2019 and 2020 . We would propose that the sentence highlighted in red is removed or amended. Ready biodegradation and inherent biodegradation inocula preparation is known to reduce the diversity of the microbial community (Martin et al. 2018). Therefore a discussion should be held with the OECD expert group to examine current scientific knowledge in terms of the recommended cellular community sizes in relation to environmental levels and under representation of microbial diversity in the closed bottle tests.

The ability to study microbial composition and diversity can be used to analyze the presence and abundance of taxa – of the overall microbial community (e.g., 16S rRNA gene amplicons), or of the active population (e.g., 16 rRNA gene transcript amplicons) – and of specific genes encoding for catabolic enzymes (metagenomics) or their expressed transcripts (metatranscriptomics). The work of Forney et al., 2001 and Goodhead et al., 2014 demonstrated that standard OECD guideline inocula preparations reduce bacterial diversity and therefore increase the variability of screening tests . Conversely, increasing cell concentrations by filtration methods led to an increase in bacterial diversity without biasing community structure and improved the probability of correctly classifying substances based on their known biodegradation behavior in screening tests (Martin et al. 2018). Further work is required away from primary degradation to ultimate but could still be considered a valuable regulatory tool.

In the light of progress and improved understanding in biodegradation in the environment we would highlight that up to 30 years have passed since the last review of the suitability of the study designs for the biodegradation tests



Agreement justifications

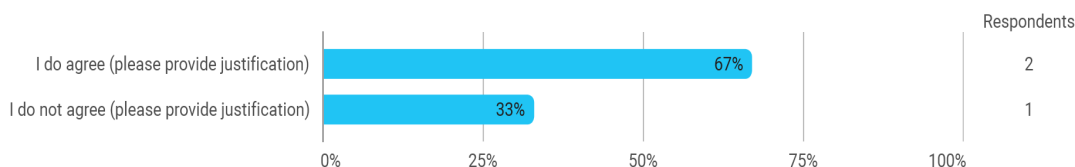
- Methods to characterize the inoculum have been developed since then. New approaches and requirements for standardization are needed.

Disagreement justifications

- It would be easier to comment if a text suggestion was proposed. New information can of course be considered. If the proposal is to use a standardized inoculum it is difficult to comment without more details.
- It is not clear what is being proposed here. If it is to use a standardized inocula, then I disagree, as it is not possible to culture everything in the environment and/or the inocula could be manipulated to give favourable results

- 5. para 32 - Measuring disappearance of the parent compound by chemical analysis does not imply mineralisation. Simulation tests are especially useful if it is known from other tests that the test substance can be mineralised and that the degradation, which is measured, covers the rate determining process. The evolution of radiolabelled CO₂ is only indicative of transformation of that labelled carbon and therefore caution should be taken in interpreting the results in the absence of non-specific analysis**

Rational: Missing information



Agreement justifications

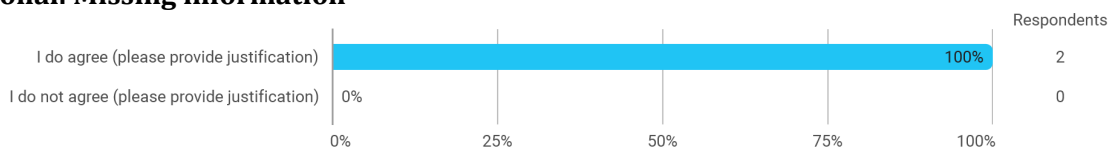
- ▶ important point
- ▶ agreed but that is common sense and well known
- ▶ It would be good to point out that just because radiolabelled CO₂ is observed, that the material may not be fully mineralized

Disagreement justifications

- ▶ There is no clear suggestion how the text should be changed. Should there be a recommendation to use specific analysis in certain cases? The comment “The evolution of radiolabelled CO₂ is only indicative of transformation of that labelled carbon and therefore caution should be taken in interpreting the results in the absence of non-specific analysis” seems to be related also to paragraph 31. If the label is in the most recalcitrant part, it can be assumed that the whole molecule is mineralized.

- 6. para 33 - Half-saturation constant; Maximum specific growth rate; These two results are not explained in the guidance texts of the simulation test guidance. If not relevant please delete or alternatively explain and include.**

Rational: Missing information



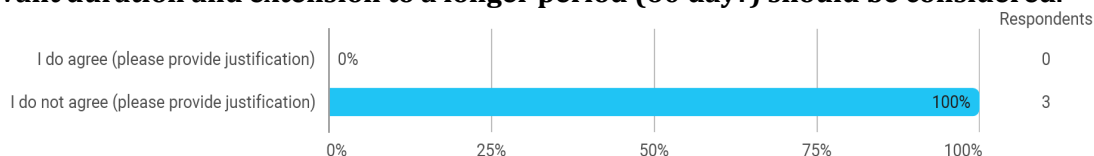
Agreement justifications

- ▶ agreed as justified above
- ▶ Agree to include further explanation or delete if not relevant.
- ▶ Please define these two terms or delete

- 7. para 41 - In order to interpret the results of a test, the full biodegradation curve should be considered so that the duration of the lag phase, slope and plateau level can be identified. The duration of 28 days in the ready biodegradability tests was defined in order to allow for sufficient time for the microorganisms to adapt to the**

chemical (lag phase) by an increase in the number of active degrading microorganisms that results in detectable degradation.

Rational: In light of the development of the enhanced tests it is questionable whether the originally 28 d test period can still be considered to reflect an environmentally relevant duration and extension to a longer period (60 day?) should be considered.

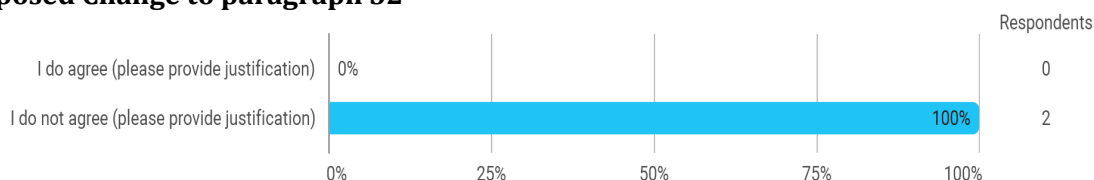


Disagreement justifications

- ready biodegradability is defined as taking place within 28 days; prolonging the duration would lead to endless discussions that some chemical is nevertheless biodegradable. It is very important to keep the requirements as they are, because the existing ready biodegradability definition is very useful to exclude substances from further assessment. A longer test duration would render the result meaningless for that purpose.
- There is no suggestion what should be changed. I don't see need to change this text.
- I do not see what the suggestion is here

8. para 52 - Amend the 'Ultimate degradation' bullet point to 'Primary degradation'

Rational: Any determined mineralisation rate has to be considered to be influenced by the position of the radiolabel and therefore this needs to be considered carefully. See Proposed Change to paragraph 32

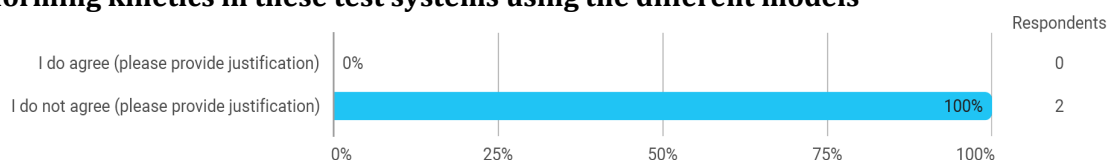


Disagreement justifications

- I don't agree to change ultimate degradation to primary degradation, because transformation products are relevant for hazard classification
- I do not think "primary" is appropriate here since it is discussing mineralization. I believe "ultimate" is correct, or would like to see more justification for why they want it changed to "primary"

9. para 57 - Amendment required: Degradation kinetics in soil or sediments may often deviate from first order kinetics because sorption/desorption processes take place simultaneously with degradation processes. Remove the following: In such cases expert judgement is required for estimating a degradation half-life or half-lives for various sub-compartment (7).

Rational: Amendment to highlight the specialist work of the FOCUS guidance for performing kinetics in these test systems using the different models

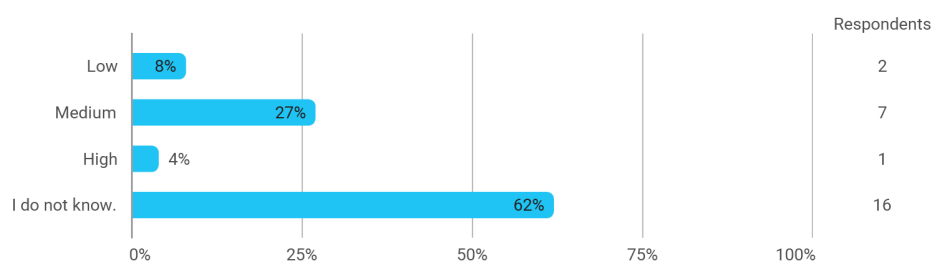


Disagreement justifications

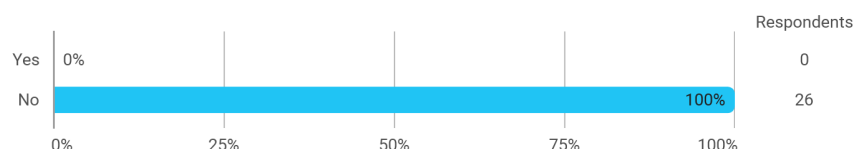
- The sentence is relevant. No justification is given for the deletion so it is difficult to comment in more detail at this point.
- I understand the intent of the proposed revision (to remove expert judgement), but something else needs to be added to the paragraph, otherwise it makes no sense.

A.14 Test No. 209: Activated Sludge, Respiration Inhibition Test (Carbon and Ammonium Oxidation)

Relevance of revision

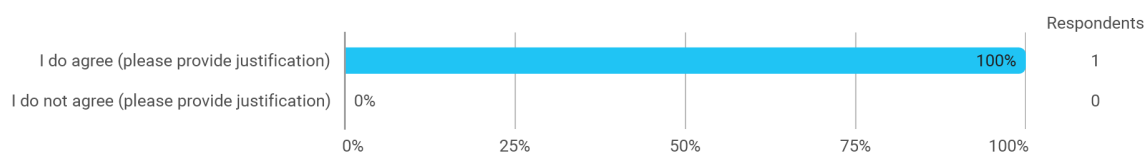


Interest in being part of an OECD Expert Group



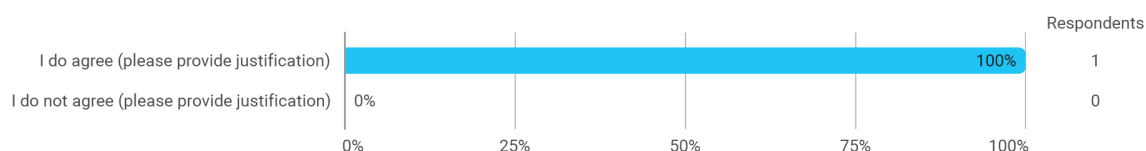
The following suggestions have been collected:

1. para 11- The blank controls (without the test substance or reference substance) oxygen uptake rate should not be less than 15 mg oxygen per one gram of activated sludge (dry weight of suspended solids) in an hour. If the rate is lower, the test should be repeated with washed activated sludge or with the sludge from another source. Rational: This is a completely new validity criterion when compared to the version before 2010. From experience we can say that the blank control activity has no relevant influence on the EC values of the reference compound 3,5-DCP. Additionally, the guideline it selves states that the quality of inoculum varies immense between different STPs. So why a threshold value for background activity. From experience we know that – even after washing and feeding overnight – an oxygen uptake rate of 20 mg O₂/(g*h) is hardly to achieve with inoculum from STPs in northern Europe. In only approx. 50% of the tests, the threshold value is surpassed. However, the effects for the reference compound and all test substances so far were comparable in runs surpassing the validity criterion and not. If there is really a scientific reason for requiring a threshold value for basic activity, it should be lowered down at least to 15 mg O₂/(g*h). This value was achieved in most of the cases (however, only after washing and feeding over night!). By the way, our source is no industrial STP. The inoculum is of high quality also in degradation studies.



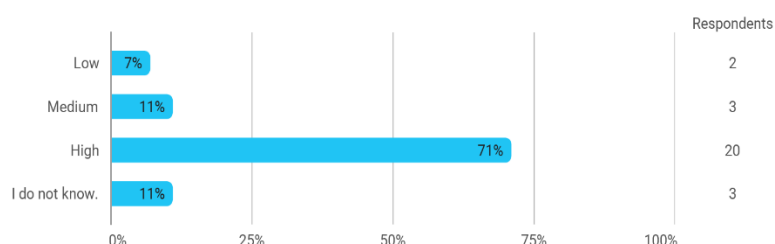
2. para 41 - The test should be carried out using a range of concentrations deduced from the preliminary test. In order to obtain both a NOEC and an ECx (e.g., EC50), six controls and five treatment concentrations in a geometric series with four replicates are in most cases recommended. In order to obtain both an ECx (e.g., EC50) only, four controls and five treatment concentrations in a geometric series with three replicates are in most cases recommended.

Rational: OECD 216 and 217 (Soil-Microorganisms) only require 3 replicates per treatment level. Even when no NOEC should be calculated, a statistical comparison of the treatment levels should be applied. At least for the EC calculation three replicates with 5 treatment levels are sufficient. Soil microorganisms and sludge microorganisms act comparable. Also, we are talking about thousands of individuals per replicate. However, even in chronic test designs with batches of animals or plants, four replicates are sufficient for NOEC and EC calculation! And in these test systems, higher variances within the replicates can be expected! For EC approaches only, the number of replicates for blank controls and treatment concentrations also can be reduced since the number of treatment levels is more relevant than the number of replicates. Since the guideline was adapted to general ecotox approaches, it is really a challenge to conduct this test design. And from experience we can say that neither the EC50, nor the EC10 is more representative after increasing the number of replicates! In contrast, the delay of the test run duration increases the variances. The inoculum biomass and activity in the stock increases in the duration of the start sequence of a sequential test design (at least after washing and feeding overnight). The respiration rate in the end blank control batch is always higher than in the start control batch! This could be due an increase of biomass in the course of the start sequence. However, it is technically not possible to measure (and adapt) the dry mass content before each application step. Therefore, for a representative result it is more important to run a test in the shortest duration possible, than to blow up the test design as much as possible.

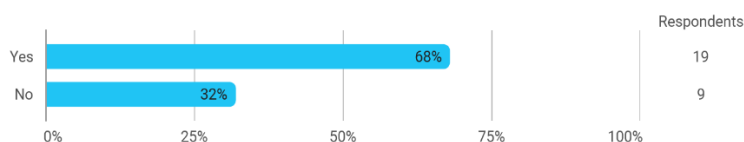


A.15 Test No. 301: Ready Biodegradability

Relevance of revision



Interest in being part of an OECD Expert Group



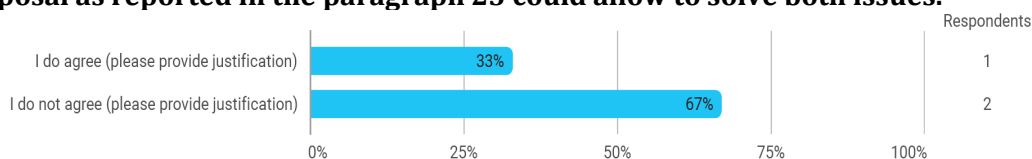
The following suggestions have been collected:

- 1. Validity of results in each Test Description (A-F) - Regarding the difference of extremes of replicate values of max. 20%: / Addition: this maximum 20% difference should be calculated as the absolute difference in percentage-points.**

Rational: In English, it is clearly indicated that the difference of extremes of replicate values should be assessed (e.g. (replicate max - replicate min)/ replicate max or min)?). However, the text in French is less clear, therefore, average values of replicates could be used based on the French guideline. I guess it is to avoid the word “réplicat”, probably considered as not really French? In this case “mesures individuelles” could be proposed. However, it should be also noted that “mesures” could indicate that we are talking of raw data.

Moreover, a clarification could be added regarding the calculation, as some laboratory proposed to compare the coefficient of variation (standard deviation / mean) to the 20% threshold. Therefore, it could be valuable to add how the “difference of extremes of replicate values” should be exactly calculated.

Of course, for most of the studies, the different calculation ways lead to the same conclusion. However, we have had already the case where the method of calculation led to different conclusions on the validity of the test. I have noted that you would like to have a clear suggestion of change, however this issue should probably be discussed to find the most appropriate way to explain this validity criteria. A calculation proposal as reported in paragraph 23 could allow to solve both issues.



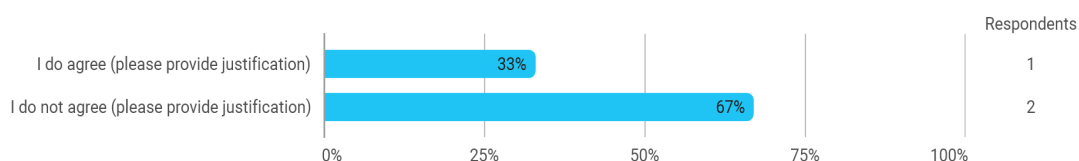
Agreement justifications

- Provides a good clarification

Disagreement justifications

- 20 pts is too high a difference. A 10-point difference already clearly indicates an issue with the test system. However, it should be clarified whether this is 20% of the lowest or the highest value. It should also be clarified what to do when more than two replicates are used.
- Caution against such a change which would affect the validity of tests - tests significantly more likely to fail. The % value of permitted variability would need to be revised. RBTs are known to be prone to high variability.

- 2. An effort should be made to harmonise test designs, e.g. harmonise inoculum source. Add specifications concerning sampling and narrow the options down to activated sludge. Review test specifications in the light of technical developments. Introduce a discussion of test limits and enhancements.**



Agreement justifications

- In general I agree, but this could negatively affect some tests (D, E) that don't use activate sludge currently. And diluting it down far enough, may cause additional complications. More lab work would be needed here.

Disagreement justifications

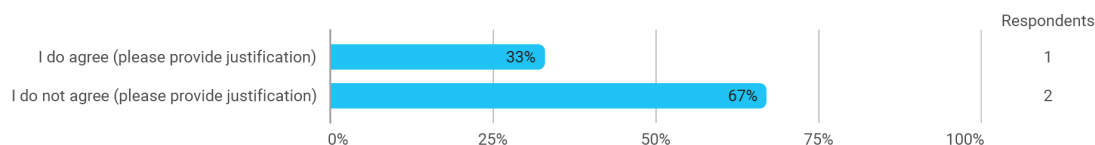
- There is no reason to narrow down to activated sludge. A discussion on test limits and enhancement could be welcome.
- Reducing the number of options is not desirable, as it reduces flexibility of the method. 301 is already a stringent test, and certain substances are problematic to deal with.

3. **para 2 - The addition of a sentence is suggested: 2. Much experience has accumulated with the six methods over the years including an OECD inter-laboratory comparison exercise (ring test) in 1988. The accumulated experience, and the ring test, have confirmed that the methods may be used for the assessment of ready biodegradability. However, depending on the physical characteristics of the substance to be tested, a particular method may be preferred. Considering the recommendation of M. Kitano and M. Takatsuki in the „Final Report of the 1988 OECD Ring-Test“* the precision of the methods described in this guideline must be determined by suitable interlaboratory test for the substance class to be investigated.**

***OECD (1988), OECD Ring-test of methods for determining ready biodegradability: Chairman's report (M. Hashimoto; MITI) and final report (M. Kitano and M. Takatsuki; CITI), Paris.**

Rational: The recommendation of M. Kitano and M. Takatsuki in the „Final Report of the 1988 OECD Ring-Test“* states, that for the determination of BOD and CO₂ evolution more reliable methods should be developed. Without further development or suitability for the application these methods have unfortunately been adopted in standards for different classes of substances, e.g. ISO 9439. This additional sentence makes it clear that it is a guideline.

***OECD (1988), OECD Ring-test of methods for determining ready biodegradability: Chairman's report (M. Hashimoto; MITI) and final report (M. Kitano and M. Takatsuki; CITI), Paris.**



Agreement justifications

- The recommendation of M. Kitano and M. Takatsuki in the „Final Report of the 1988 OECD Ring-Test“* states, that for the determination of BOD and CO₂ evolution more reliable methods should be developed. Without further development or suitability for the application

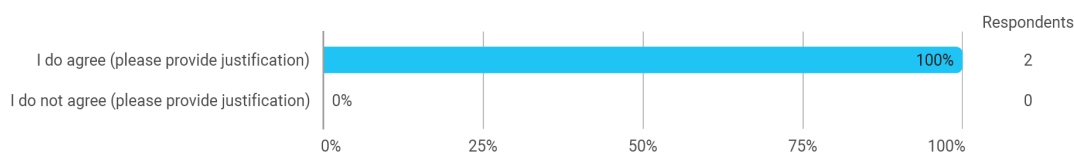
these methods have unfortunately been adopted in standards for different classes of substances, e.g. ISO 9439. This additional sentence makes it clear that it is a guideline.

Disagreement justifications

- It seems like a lot of unnecessary work to do lots of interlaboratory tests on many chemicals to establish the applicability of a specific test.
- Neither agree nor disagree: We don't understand the aim of this addendum

4. para 2 - Acknowledgment that the proposed methods are not equivocal in their performance and their outcome is required.

Rational: There are seven different types of test designs, which are widely used, and inocula can be selected from five different environmental sources with different concentrations and microbial communities (OECD 1992; OECD 2014; Kowalczyk et al. 2015). The inocula can further be prepared in up to four different ways (Goodhead et al., 2014). Consequently, RBT outcomes for a given substance can differ widely (Thouand et al. 1995; Vazquez-Rodriguez et al. 2003; Goodhead et al. 2014). This does imply that an unequivocal result is achievable.



Agreement justifications

- A brief acknowledgement would be useful
- Propose to include the rationale along with this suggestion, as follows: There are seven different types of test designs, which are widely used, and inocula can be selected from five different environmental sources with different concentrations and microbial communities (OECD 1992; OECD 2014; Kowalczyk et al. 2015). The inocula can further be prepared in up to four different ways (Goodhead et al., 2014). Consequently, RBT outcomes for a given substance can differ widely (Thouand et al. 1995; Vazquez-Rodriguez et al. 2003; Goodhead et al. 2014). This does imply that an unequivocal result is achievable.

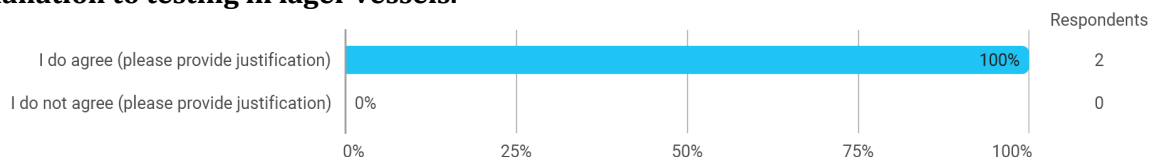
5. para 6 - Original paragraph: Normally, the test lasts for 28 days. Tests however may be ended before 28 days, i.e. as soon as the biodegradation curve has reached a plateau for at least three determinations. Tests may also be prolonged beyond 28 days when the curve shows that biodegradation has started but that the plateau has not been reached by day 28, but in such cases the chemical would not be classed as readily biodegradable. Change as proposed in April 2005: 21. The standardised test duration is 28 days although tests may be prolonged beyond 28 days if the biodegradation has started but has not yet reached a plateau. However, only the extent of biodegradation achieved within 28 days should be used for the evaluation of ready biodegradability but degradation after 28 days would allow the test substance to be classified as inherently biodegradable (see paragraph 36)

Rational: The explanation was given in Paragraph 36 of the proposed introduction from April 2005:

Since inherent biodegradability can be considered to be a specific property of a chemical, it is not necessary to define limits on test duration or biodegradation rates. Biodegradation above 20% of theoretical (measured as BOD, DOC removal or COD)

may be regarded as evidence of inherent, primary biodegradability, whereas biodegradation above 70% of theoretical (measured as BOD, DOC removal or COD) may be regarded as evidence of inherent, ultimate biodegradability. When results of ready biodegradability tests indicate that the pass level criterion is almost fulfilled (i.e. ThOD or DOC slightly below 60% or 70% respectively) such results can be used to indicate inherent biodegradability. This is also the case when the pass level criterion is fulfilled but the 10-day window criterion is not. Such application of ready biodegradability tests, which may include their incubation beyond 28 days, may in some cases eliminate the need for additional testing of biodegradability in inherent or simulation tests.

Explanation to testing in larger vessels.

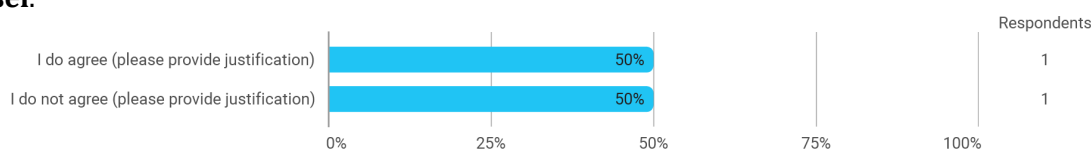


Agreement justifications

- This would be a good clarification
- We agree and would suggest increasing environmental relevance of test.

6. further suggestion: Testing in larger vessels and therefore increasing e.g. the volume of inoculum to increase microbial diversity.

Rational: At very small test volumes the total number of and the number of different types of microorganisms introduced into the test flask decreases. Conducting biodegradation tests using larger volumes of environmental waters increases the total number of microorganisms introduced into the test, and the number of different types, without changing the density of microorganisms introduced (Ingerslev et al., 2001). This will increase the probability of introducing a competent microorganism into the test vessel.



Agreement justifications

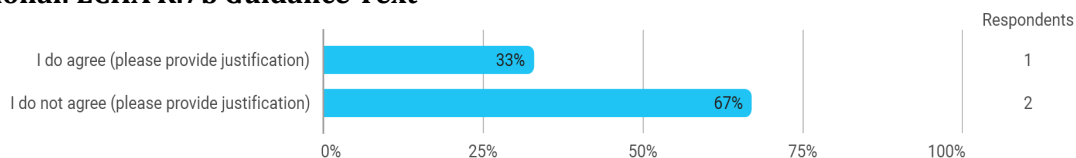
- In theory this should work to increase diversity and decrease the variability of the tests. additional lab work may be required.

Disagreement justifications

- Neither agree nor disagree: It is not clear to which paragraph this refers and what is its justification

7. para 6 - This paragraph needs to be updated to include extension to the test duration up to sixty days can be accepted

Rational: ECHA R.7b Guidance Text



Agreement justifications

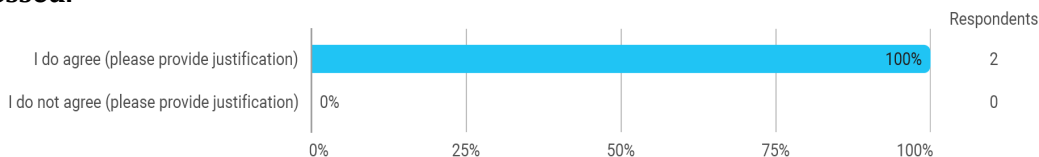
- We agree. Many tests have been run until 60 days instead of 56.

Disagreement justifications

- This comment needs more context. Are they wanting to extend a test to 60 days and still claim ready biodegradability? It is not clear.
- 60 day test is not considered a ready biodegradation test.

- para 9 - As an RBT is key data for many substance registrations, further development and guidance are needed to improve ready biodegradability tests of difficult to test substances. Recent academic publications have highlighted advances in passive dosing of hydrophobic substances to these test systems with positive results.**

Rational: Difficulties encountered in estimating the biodegradability of poorly water-soluble substances are often linked to their aqueous solubility and limited bioavailability to microorganisms (Stucki and Alexander 1987; Alexander 1999). Laboratory tests according to OECD or International Organization for Standardization (ISO) guidelines prescribe, in many cases, test substance concentrations well above the solubility limit for poorly soluble substances, since testing in the ng/L to µg/L range is experimentally challenging for many approaches (Sweetlove 2017). However, operating biodegradation tests near or above the solubility limit can lead to an underestimation of biodegradability when dissolution of the chemical becomes rate limiting or when high test concentrations inhibit the biodegradation process (Hammershøj et al. 2019; Hammershøj et al. 2020). These results were confirmed when longer half-lives were observed for hydrophobic substances exposed via passive dosing when higher test concentrations approaching their water solubility were assessed.



Agreement justifications

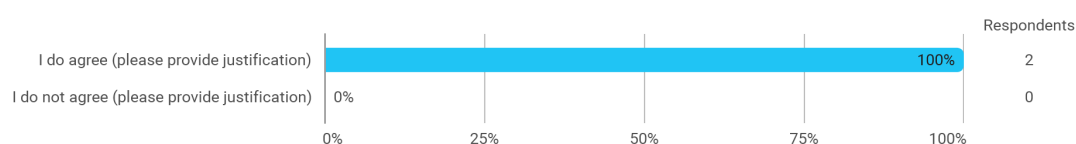
- More work is needed to update the guidance for how to handle poorly soluble substances
- We agree and suggest including this in an appendix of the TG and not merely in the guidance.

- para 17 - This text and associated paragraphs does not reflect how conservative the inocula density and community size are within these studies. It is recommended to present this in the revision of this test guideline.**

Rational: It is questionable whether the relatively small sample sizes used for biodegradation tests are sufficient to represent the microbial community encountered by substances freely diffusing throughout a given environment. Additionally the variability in the results indicate that this needs to be addressed.

Inoculum density in an ready biodegradability test (RBT) is ~ 104 cells/mL, this is typically ten thousand times less concentrated than the same community in a typical activated sludge wastewater treatment plant (WWTP) (108 cells/mL), and the total number of cells in an RBT test vessel, 108, is a billion times fewer than that of a small sized WWTP (~1017 in 100 m3 reactor), with ensuing differences in diversity. Across different biodegradation test systems, the cell concentrations (101 cells/mL in some RBTs to 1010 cells/g in soil simulation tests) and total number of microorganisms vary widely (105 to 1010).

Additional thought should be given to the academic work of both Martin et al., 2018 and Ott et al., 2020 where methodology was presented that would allow for concentration of an innoculum to environmentally relevant concentration but also reported improved methodology in determining the cell concentration and characterisation of the inocula used.

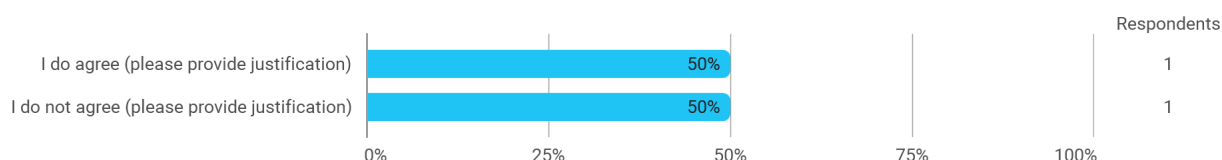


Agreement justifications

- Some additional contexts on the conservative nature of the inocula is needed
- It is important that the TG becomes clearer that a negative result does not mean that the substance is not readily biodegradable (in other words, the test is so stringent that a positive results should always supersede negative results). This would require a modification of §24

10. Table 2 - It should be clarified in Table 2 that the test conditions specified in relation to inoculum density are cumulative and not alternative measures on inoculum density

Rational: There seems to be a misunderstanding that the concentration in suspended solids is not to be regarded as a valid measure of inoculum density. When sludge is used the inoculum should meet both the criteria in terms of mg/l SS and approx. cells/ml. For an effluent, both the conditions in terms of volume of added effluent and cell density should be met. A suspended solid concentration or volume of effluent does not provide a direct estimate of bacterial density and can be subject to various biases. This should be clarified in the TG.

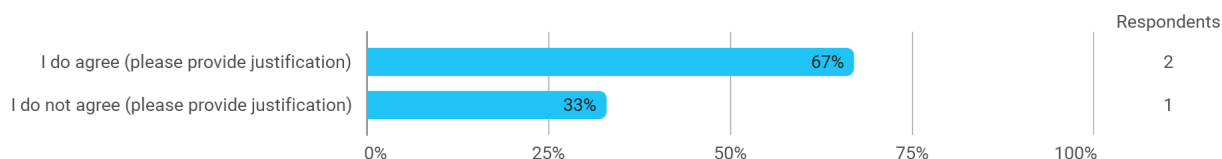


Disagreement justifications

- In practice, inoculum in cell density per L is never measured. Adding this consideration would invalidate almost 100% of the tests conducted so far. Plus, the added value of doing such a thing is unclear.

11. ANNEX 3: It is not recommended that solid carriers be used for solid test substances but they may be suitable for oily substances. However, solid test substances can be dissolved in volatile organic solvents and applied on solvent carriers (e.g. silica gel) to increase the bioavailability of the test substance.

Rational: Silica gel is a solid carrier, listed in OECD 310. However, since it is not listed in OECD 301 as an allowed carrier, authorities already refused valid OECD 301 tests - even when additional blank controls with only the carrier were applied. It should be more apparant, that any utility listed in one Screening Degradation Guideline should be allowed in any other Screening Degradation Study according to any other guideline. This is true especially for inoculum sources and test item carriers.



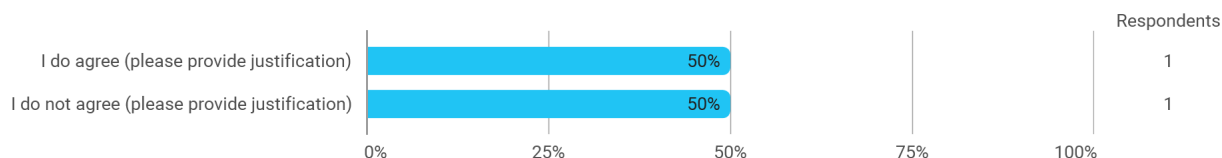
Agreement justifications

- It is unclear what exactly is meant with this suggestion. If it refers to using passive dosing for poorly soluble substances, though, we agree.
- Suggest aligning Annex 3 with ISO 10634 (2018)

Disagreement justifications

- I agree that more work is needed for poorly soluble substances. However, dissolving solid substances and coating them on silica gel, may not be the appropriate method, as these solid substances could form a film on the silica gel, thus reducing the bioavailability of the materials

12. It needs to be clarified that in the Table 2 information on inoculum concentration lists the parameters which all (in general) should be respected, but not only any one of them, the most important being information on the number of active cells per litre or millilitre. The same applies to OECD TG 310.



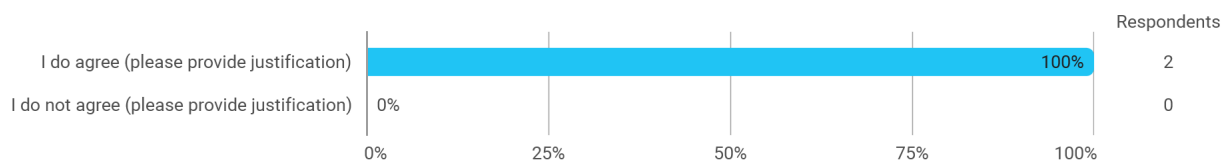
Agreement justifications

- I agree, clarification needed

Disagreement justifications

- In practice, inoculum in cell density per L is never measured. Adding this consideration would invalidate almost 100% of the tests conducted so far. Plus, the added value of doing such a thing is unclear.

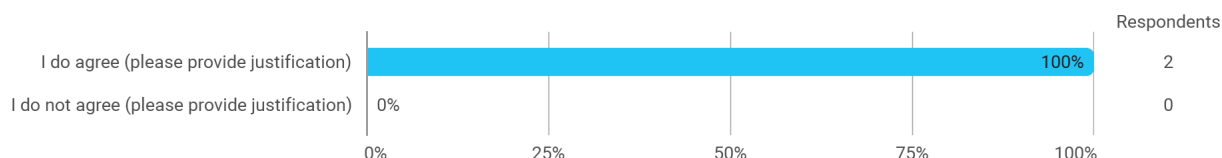
13. Include a 60 day extension to tests 301 and 310 – if the ingredient then passes the threshold, the screening test could be used to formally classify ingredients as “inherently and ultimately” biodegradable.



Agreement justifications

- This is similar to other suggestions. Clarification is needed to allow extension to 60 days to allow for classification of inherent ultimate and primary
- We agree if this means 60% threshold.

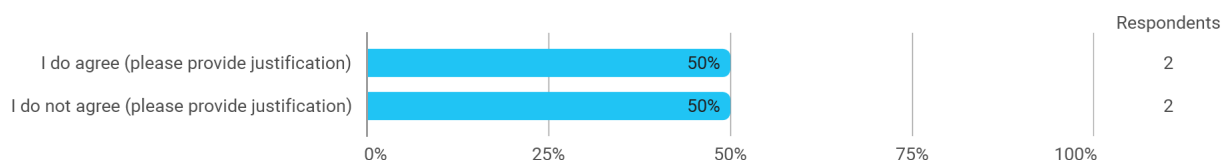
14. In tests 301, 310 & 306 – increase the biomass to increase the likelihood of degraders being present. Consider potential for inclusion of a series of vessels in study design with increased biomass (alongside standard biomass) to reduce the biological lottery of not having competent degraders present.



Agreement justifications

- Laboratory verification work would be needed, but in general more microorganisms
- We agree, although not for the exact same reasons. We think it shouldn't be about finding a competent degrader by all means, but rather about improving the reliability and the reproducibility of the test. It is a fact that most substance failing a RBT are yet not persistent when tested in a simulation test. This triggers too much unnecessary testing. One option is to increase the biomass to an acceptable level.

15. Allow the use of adapted inocula in the 301 and 310 tests.



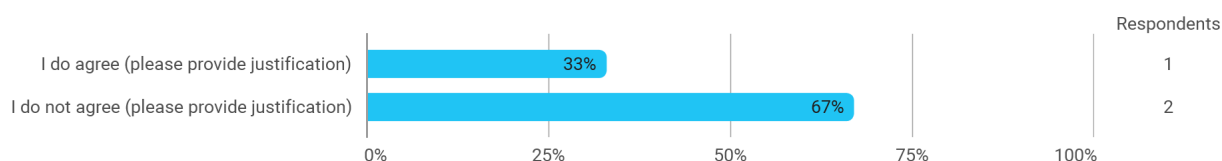
Agreement justifications

- Moderate adaptation should be acceptable as an add-on to the 301 test, A number of chemicals have shown to adapt fairly quickly after commercialisation and this can be simulated using appropriate conditions in the laboratory.
- We agree, if the purpose it to increase reproducibility and relevance.

Disagreement justifications

- This will require some significant discussion and/or lab work before it can be accepted, as it is a pretty significant shift in the design of the tests.
- Pre-adapted inocula would not be appropriate for this test, which is intended to be stringent.

16. Reduce the number of 301 tests to only those which are unequivocal of mineralisation. Remove 301a and 301e.



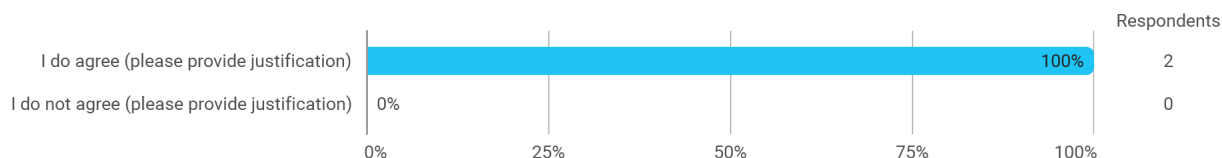
Agreement justifications

- DOC can be ambiguous in its interpretation. Given the availability of CO₂ and O₂ endpoints, DOC is not needed as an endpoint.

Disagreement justifications

- More discussion is needed before removing the two tests based on DOC, but I may be supportive of the proposal since they offer now clear advantages over some of the other guidelines
- Removing those tests would invalidate all OECD 301A and OECD 301E performed so far.

17. Additional reference compounds in both screening (301 series, 302) and simulations (307,308,309) tests should be included. These additional reference compounds should allow to reflect degradability of more complex compounds ((bio)polymers).



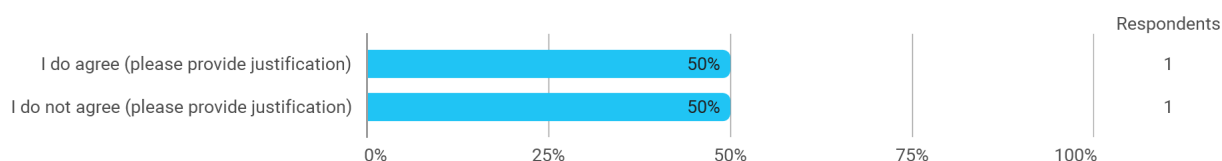
Agreement justifications

- Agree, guidance is needed on how to handle polymers, including an appropriate positive control. We have had good results with dextran
- We fully agree. Benzoate as a reference is not a reliable reference, as even tap water can mineralize benzoate.

18. para 2 - Include a phrase that the outcomes of different OECD 301 test methods are not comparable to each other, depending on the test substance. Rephrase the last sentence to read "However, depending on the physical characteristics of the substance to be tested, a particular method may be required to obtain a valid result).

Rational: The recommendation of M. Kitano and M. Takatsuki in the „Final Report of the 1988 OECD Ring-Test“* states, that for the determination of BOD and CO₂ evolution more reliable methods should be developed. Without further development or suitability for the application these methods have unfortunately been adopted in standards for different classes of substances, e.g. ISO 9439. This additional sentence makes it clear that it is a guideline.

***OECD (1988), OECD Ring-test of methods for determining ready biodegradability: Chairman's report (M. Hashimoto; MITI) and final report (M. Kitano and M. Takatsuki; CITI), Paris.**



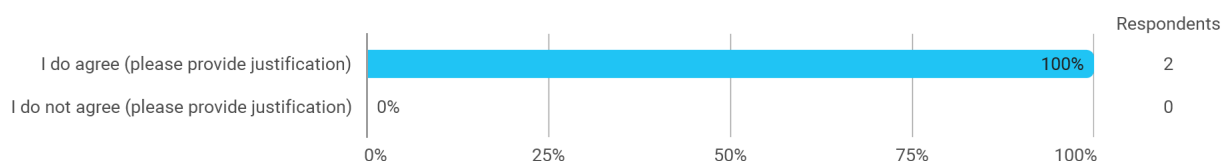
Agreement justifications

- We fully agree to make §2 clearer.

Disagreement justifications

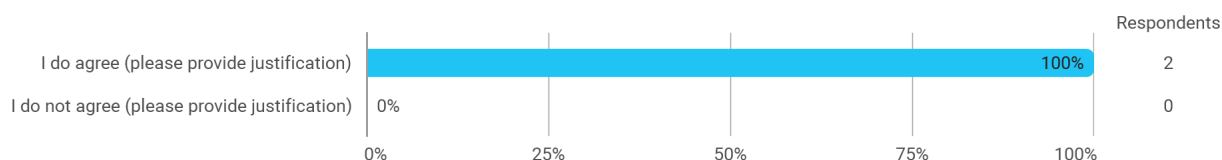
- This is a slippery slope. who gets to determine which test method is valid for which material? Could lead to some prolonged disagreements. More clarification will be needed than is provided in the suggestion.

19. para 10 - Currently, the guideline states "... These pass values have to be reached in a 10-d window within the 28-d period of the test, except where mentioned below". It is not clear where "below" is in the guidelines. Suggest removing "below and stating explicitly " These pass values have to be reached in a 10-d window within the 28-d period of the test, except for in the MITI method and for mixtures."



- Agreement justifications
- Agree, explicitly state the exceptions in this paragraph
- We agree and would specify further "except in the MITI method and for multiconstituents and UVCB substances and for mixtures".

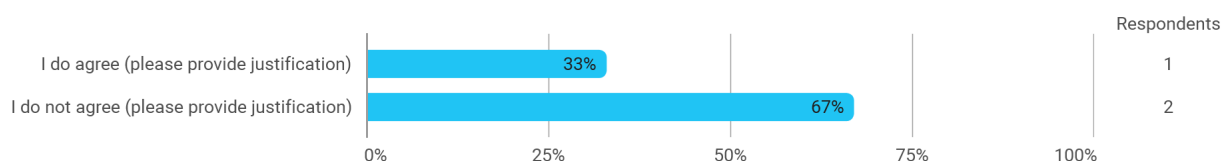
20. para 11 - Include a statement like "The inclusion of a reference compound allows verification of the microbial activity in the inoculum. Suitable compounds would be those that are widely known to be readily biodegradable and, optionally, may bear similarity to the test compound."



Agreement justifications

- I agree with premise, but I think a bit more context is needed than is provided in the suggestion
- We agree that more relevant reference compounds need to be proposed by the TG.

21. para 12 - Suggest changing "determinations should be carried out at least in duplicate" to "determinations where the outcome is fairly certain should be carried out at least in duplicate, but to improve the reliability of the test, it is recommended to perform the test at least in triplicate"



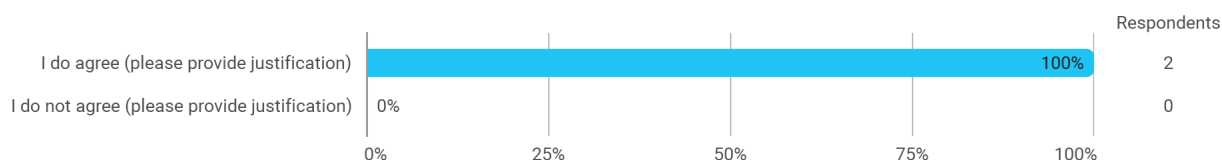
Agreement justifications

- We agree that the TG should allow more space to increase the number of replicates. However, the choice of the number should not be based on intuition. It would be challenging to predict that the outcome is "fairly certain"

Disagreement justifications

- I'm not sure this is needed. Perhaps something less strongly worded, like "determinations may be carried out in triplicate to improve the reliability of the test". Regardless, they would still need to be within 20%, which also applies to the duplicate determinations, so I'm not sure how useful this is.
- Not clear how it can be determined whether the outcome will be 'fairly certain'. Mandating additional vessels will affect how likely studies are to fail validity criteria for replicate variability (although will increase robustness of results).

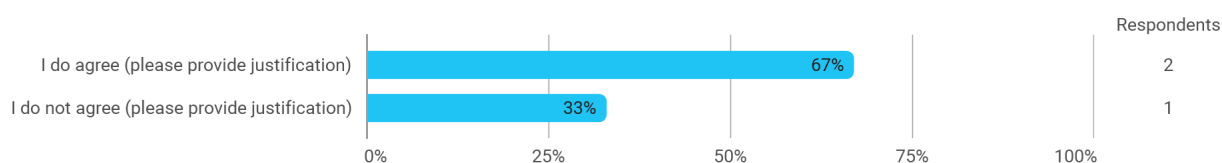
22. para 12 - Include a discussion on bioavailability, its effects on the variability of test results, and suggest methods by which experiments can be enhanced to improve upon this limitation (integrate knowledge from ISO 10634 guideline; Sweetlove et al. 2016; and other pertinent studies)



Agreement justifications

- Agree, more info is needed on bioavailability and how to test poorly soluble substances
- We agree that the bioavailability issue is fundamental. The annexes of the TG should be enriched to better deal with this matter and allow people to avoid performing useless simulation tests

23. para 12 - Provide guidance on the number of bioavailability improvement methods (BIMs) which are allowable in a single test



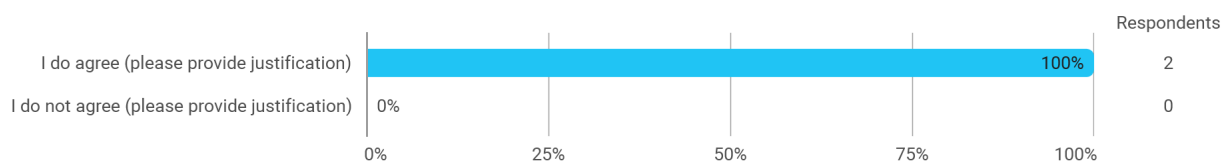
Agreement justifications

- Agree, more info is needed on bioavailability and how to test poorly soluble substances
- Fully agree. Recommendation could be included in an annex.

Disagreement justifications

- Don't see it necessary to put a limit on this.

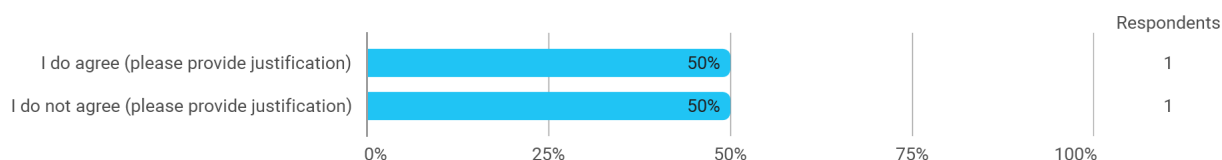
24. para 12 - Discuss and suggest a low-solubility control (reference substance) to determine a system's function following implementation of bioavailability improvement methods (BIMs)



Agreement justifications

- Agree, more info is needed on bioavailability and how to test poorly soluble substances
- We fully agree that more relevant reference controls should be allowed in the TG.

25. paras 17/18 - Provide greater guidance on the quality and characterization of the inoculum. Inoculum quality refers to the number of microbes, the potential for competent degraders and biochemical activity



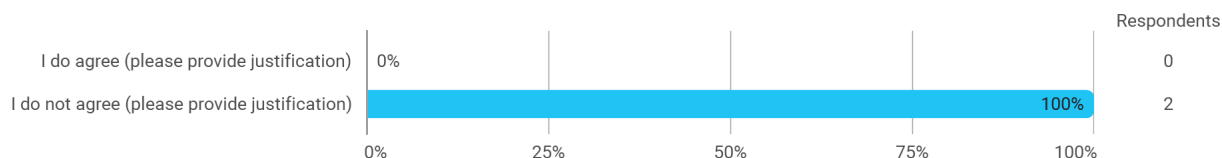
Agreement justifications

- More guidance is needed on inoculum quality, but I don't think it should be a validity criteria

Disagreement justifications

- The TG is made to be simple and cheap. Inoculum characterization may lead to additional burden for little added value.

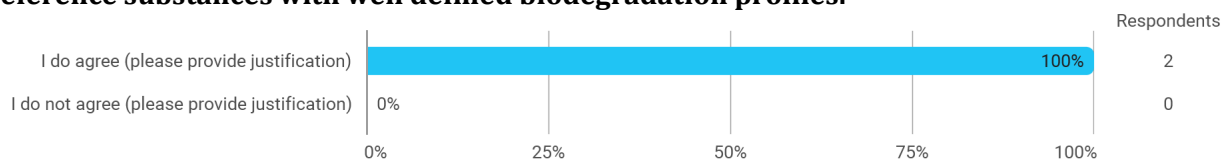
26. paras 17/18 - Review and amend the suggested microbial density in the test protocol and include the microbial load (number of microbes) as a test validity criteria.



Disagreement justifications

- This adds additional burden on testing labs, and which method of determining density/diversity is appropriate?
- We think that this provides no added value, further burdens CROs' workload and these data have been rarely recorded so far. Adding this as a validity criterion would invalidate all studies conducted so far.

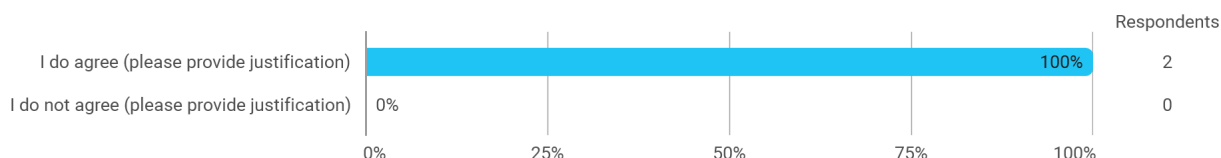
27. paras 17/18 - Provide benchmarking options/tools in the form of a range of reference substances with well defined biodegradation profiles.



Agreement justifications

- Info on more reference substances is needed

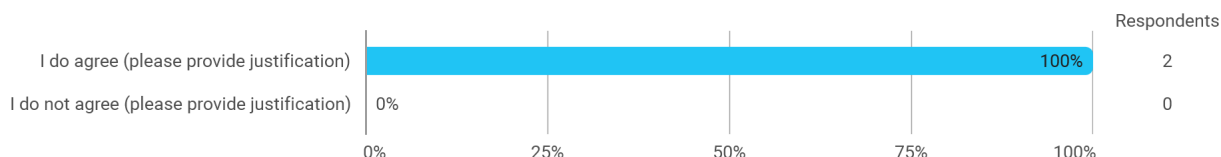
28. paras 17/18 - Further guidance on the term 'adaptation' of microbes is required. What is the official definition, and how is it possible to measure/conclude as much.



Agreement justifications

- More discussion is needed on if adaptation is allowed, and if so, what are the allowable parameters.
- Good question !

29. Annex III - The guideline could better describe the methods that can be used for dealing with difficult substances (type of emulsifier, solvent or carrier to be used, dosage, for which kind of substance properties...)

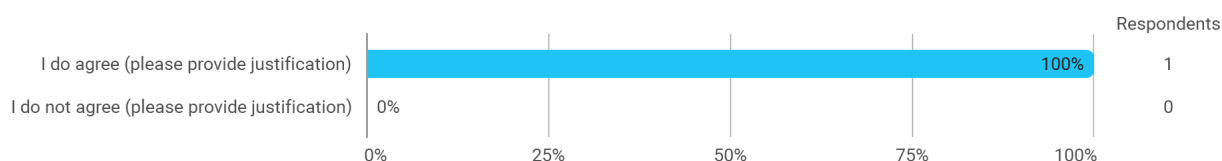


Agreement justifications

- Agree, more info is needed on bioavailability and how to test poorly soluble substances
- Fully agree !

30. para 2 - Acknowledgment that the proposed methods are not equivocal in their performance and their outcome is required.

Rational: There are seven different types of test designs, which are widely used, and inocula can be selected from five different environmental sources with different concentrations and microbial communities (OECD 1992; OECD 2014; Kowalczyk et al. 2015). The inocula can further be prepared in up to four different ways (Goodhead et al., 2014). Consequently, RBT outcomes for a given substance can differ widely (Thouand et al. 1995; Vazquez-Rodriguez et al. 2003; Goodhead et al. 2014). This does imply that an unequivocal result is achievable.

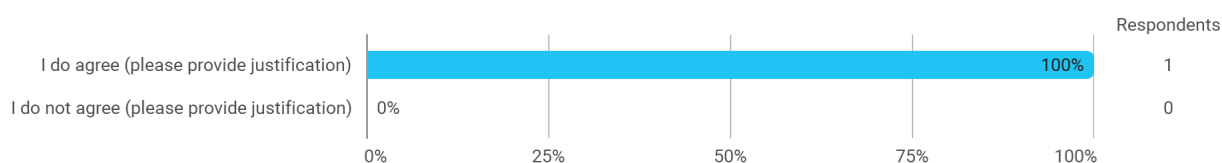


Agreement justifications

- A brief acknowledgement would be useful

31. para 6 - This paragraph needs to be updated to include extension to the test duration up to sixty days can be accepted

Rational: ECHA R.7b Guidance Text

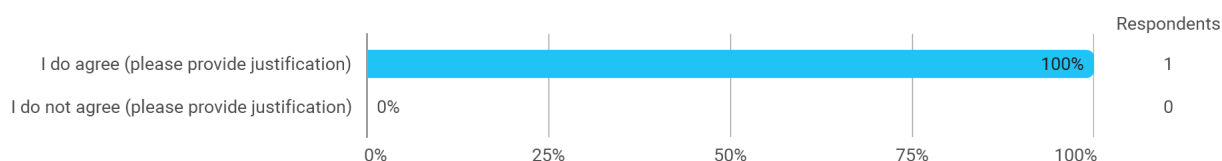


Agreement justifications

- Similar to other suggestions, allow up to 60 days for determining inherent ultimate and primary degradation

32. para 9 - As an RBT is key data for many substance registrations further development and guidance are needed to improve ready biodegradability tests of difficult to test substances. Recent academic publications have highlighted advances in passive dosing of hydrophobic substances to these test systems with positive results

Rational: Difficulties encountered in estimating the biodegradability of poorly water-soluble substances are often linked to their aqueous solubility and limited bioavailability to microorganisms (Stucki and Alexander 1987; Alexander 1999). Laboratory tests according to OECD or International Organization for Standardization (ISO) guidelines prescribe, in many cases, test substance concentrations well above the solubility limit for poorly soluble substances, since testing in the ng/L to µg/L range is experimentally challenging for many approaches (Sweetlove 2017). However, operating biodegradation tests near or above the solubility limit can lead to an underestimation of biodegradability when dissolution of the chemical becomes rate limiting or when high test concentrations inhibit the biodegradation process (Hammershøj et al. 2019; Hammershøj et al. 2020). These results were confirmed when longer half-lives were observed for hydrophobic substances exposed via passive dosing when higher test concentrations approaching their water solubility were assessed.



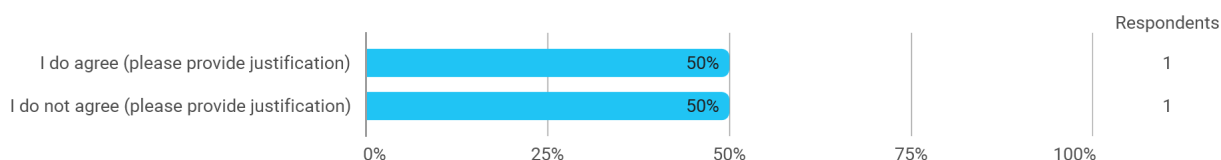
Agreement justifications

- More work is needed to update the guidance for how to handle poorly soluble substances

33. para 17 - This text and associated paragraphs does not reflect how conservative the inocula density and community size are within these studies. We would recommend that this is presented in the revision of this introduction and any proposed updates in the test guidelines. In all the tests, the test substance providing the sole source of organic carbon (except for carbon associated with the biomass) is diluted in a test medium containing a relatively low concentration of biomass. In all the tests, a non-specific analytical method is used to follow the course of biodegradation.

Rational: It is questionable whether the relatively small sample sizes used for biodegradation tests are sufficient to represent the microbial community encountered by substances freely diffusing throughout a given environment. Additionally the variability in the results indicate that this needs to be addressed. Inoculum density in an ready biodegradability test (RBT) is ~ 104 cells/mL, this is typically ten thousand times less concentrated than the same community in a typical activated sludge wastewater treatment plant (WWTP) (108 cells/mL), and the total number of cells in an RBT test vessel, 108, is a billion times fewer than that of a small sized WWTP (~1017 in 100 m3 reactor), with ensuing differences in diversity. Across different biodegradation test systems, the cell concentrations (101 cells/mL in some RBTs to 1010 cells/g in soil simulation tests) and total number of microorganisms vary widely (105 to 1010).

Additional thought should be given to the academic work of both Martin et al., 2018 and Ott et al., 2020 where methodology was presented that would allow for concentration of an innoculum to environmentally relevant concentration but also reported improved methodology in determining the cell concentration and characterisation of the inocula used.



Agreement justifications

- Some additional context on the conservative nature of the inocula is needed

Disagreement justifications

- Neither agree nor disagree: We are not sure we understand the reason for this. The stringency aspect of these tests is already stated in the TG.

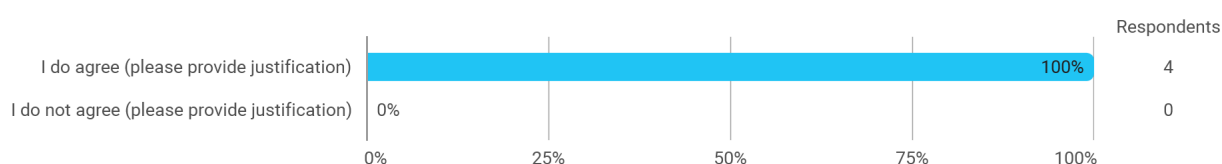
34. para 24 - Difference of extremes of replicate values should be defined more explicitly. E.g. is this measured at the plateau, end of test, or end of 10 day window. Is the study director free to choose? Is the test invalid if this substance fails at any one of these time points? What happens when tests utilise more replicates, thereby increasing statistical power, but increasing the probability of failing this criterion? Should variation be expressed in a different way (e.g. as % coefficient of variation)? As an absolute figure of 20% is relatively much larger in cases of low biodegradation, compared with high biodegradation.

Rational: In English, it is clearly indicated that the difference of extremes of replicate values should be assessed (e.g. (replicate max - replicate min)/ replicate max or min?). However, the text in French is less clear, therefore, average values of

replicates could be used based on the French guideline. I guess it is to avoid the word “réplicat”, probably considered as not really French? In this case “mesures individuelles” could be proposed. However, it should be also noted that “mesures” could indicate that we are talking of raw data.

Moreover, a clarification could be added regarding the calculation, as some laboratory proposed to compare the coefficient of variation (standard deviation / mean) to the 20% threshold. Therefore, it could be valuable to add how the “difference of extremes of replicate values” should be exactly calculated.

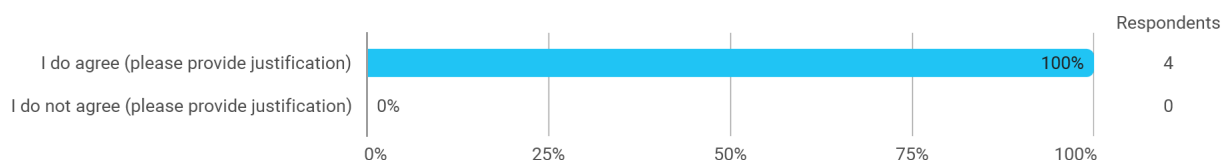
Of course, for most of the studies, the different calculation ways lead to the same conclusion. However, we have had already the case where the method of calculation lead to different conclusion on the validity of the test. I have noted that you would like to have a clear suggestion of change, however this issue should probably be discussed to find the most appropriate way to explain this validity criteria. A calculation proposal as reported in the paragraph 23 could allow to solve both issues.



Agreement justifications

- Current validity criteria are ambiguous. Adjusting the way that variation between replicates is determined must be done with care to not impact the performance of the test.
- More clarification is needed on the % difference in replicates for validity criteria, but additional work would be needed if we want to switch to different metrics
- Very relevant questions. The TG should be clarified.
- Current validity criteria are ambiguous. Adjusting the way that variation between replicates is determined must be done with care to not impact the performance of the test.

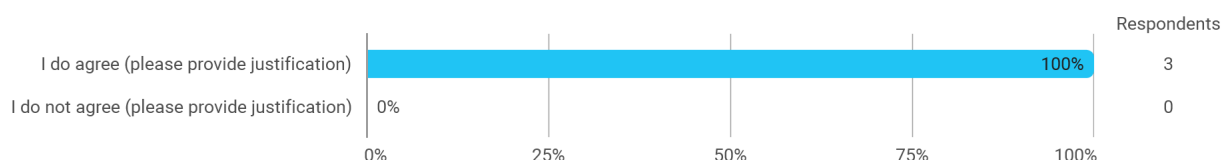
35. table 2 - Clarify that for manometric respirometry test both 100 mg test substance /L and 50-100 mg ThOD/L test concentrations are acceptable. Many substances have a ThOD > 1 mg O₂ / mg test substance, hence 50-100 mg ThOD/L is typically << 100 mg test substance /L



Agreement justifications

- Current guideline text is ambiguous and this would help to resolve this.
- This clarification is needed
- In addition, OECD 301F should allow reducing the initial concentration of test substance up to 1-3 mg/L in case of poorly soluble substance (< 1 mg/L). From experience, this works quite well.
- Current guideline text is ambiguous and this would help to resolve this.

36. Annex III - Include reference to latest ISO 10634 standard, where additional work has been done concerning bioavailability improvement methods.

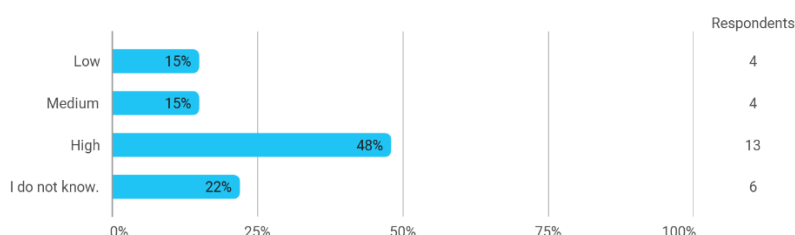


Agreement justifications

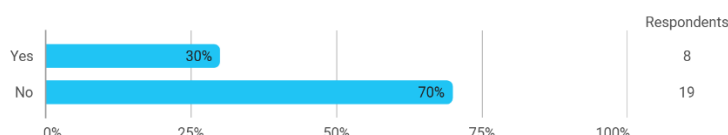
- update to correspond to latest ISO standard
- Agree, more info is needed on bioavailability and how to test poorly soluble substances
- Very relevant proposal.

A.16 Test No. 306: Biodegradability in Seawater

Relevance of revision

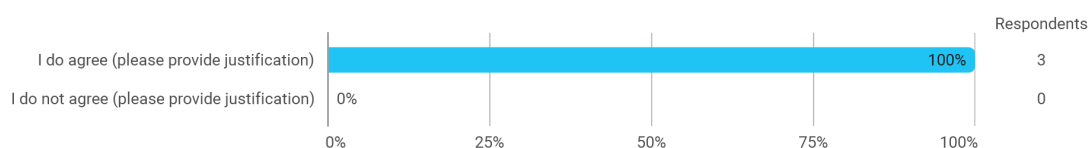


Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. In tests 301, 310 & 306 – increase the biomass to increase the likelihood of degraders being present. Consider potential for inclusion of a series of vessels in study design with increased biomass (alongside standard biomass) to reduce the biological lottery of not having competent degraders present.
Update 306 to include CEFIC ring test of cell concentrations ECO11.
See paper: Ott, A., Martin, T.J., Acharya, K., Lyon, D.Y., Robinson, N., Rowles, B., Snape, J.R., Still, I., Whale, G.F., Albright, V.C., Bäverbäck, P., Best, N., Commander, R., Eickhoff, C., Finn, S., Hidding, B., Maischak, H., Sowders, K.A., Taruki, M., Walton, H.E., Wennberg, A.C., Davenport, R.J., 2020. Multi-laboratory Validation of a New Marine Biodegradation Screening Test for Chemical Persistence Assessment. Environ. Sci. Technol. 54, 4210–4220. <https://doi.org/10.1021/acs.est.9b07710>

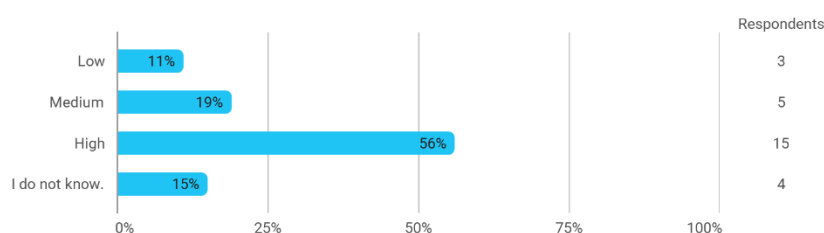


Agreement justifications

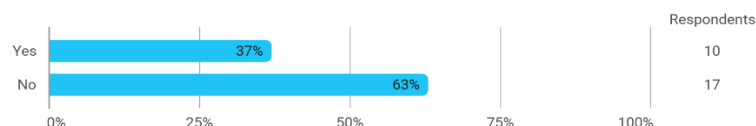
- inoculum diversity in OECD 306 is currently not representative
- I agree that increased inocula concentrations provide better results. I know that refinements based on the paper listed are already making their way through the system, but I don't know if this will revise the 306, or produce a new guideline
- We agree, although not for the exact same reasons. We think it shouldn't be about finding a competent degrader by all means, but rather about improving the reliability and the reproducibility of the test. It is a fact that most substance failing a RBT are yet not persistent when tested in a simulation test. This triggers too much unnecessary testing. One option is to increase the biomass to an acceptable level.

A.17 Test No. 310: Ready Biodegradability - CO₂ in sealed vessels (Headspace Test)

Relevance of revision



Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. **para 26 - It should be clarified that adequate inoculum density needs to be demonstrated based on the number of colony-forming units per millilitre in the final mixture. This also applied to Table 2 in OECD TG 301**

Rational: There seems to be a misunderstanding that the concentration in suspended solids is not to be regarded as a valid measure of inoculum density. When sludge is used the inoculum should meet both the criteria in terms of mg/l SS and approx. cells/ml. For an effluent, both the conditions in terms of volume of added effluent and cell density should be met. A suspended solid concentration or volume of effluent does not provide a direct estimate of bacterial density and can be subject to various biases.



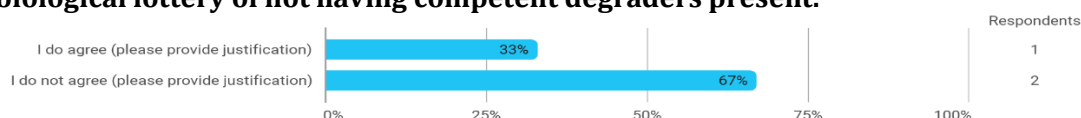
Agreement justifications

- It may be fine to demonstrate microbial density, but I don't think it should be a validity criteria without any additional work.

Disagreement justifications

- In practice, inoculum in cell density per L is never measured. Adding this consideration would invalidate almost 100% of the tests conducted so far. Plus, the added value of doing such a thing is unclear.

2. In tests 301, 310 & 306 – increase the biomass to increase the likelihood of degraders being present. Consider potential for inclusion of a series of vessels in study design with increased biomass (alongside standard biomass) to reduce the biological lottery of not having competent degraders present.



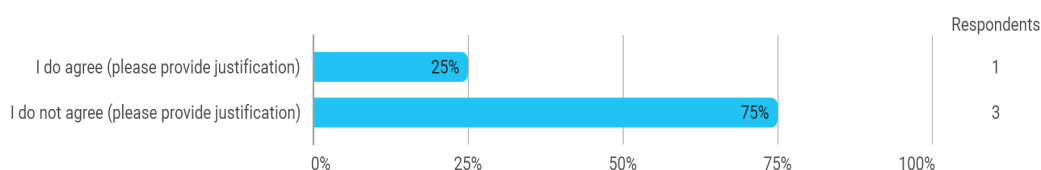
Agreement justifications

- We agree, although not for the exact same reasons. We think it shouldn't be about finding a competent degrader by all means, but rather about improving the reliability and the reproducibility of the test. It is a fact that most substance failing a RBT are yet not persistent when tested in a simulation test. This triggers too much unnecessary testing. One option is to increase the biomass to an acceptable level.

Disagreement justifications

- This would change the test rational, outcome and usability
- More work and discussion is needed before the use of increased inoculum concentrations should be allowed

3. Allow the use of adapted inocula in the 301 and 310 tests.



Agreement justifications

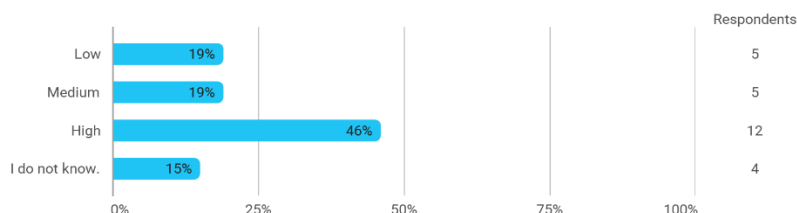
- We don't disagree, although we think the purpose of this should be stated.

Disagreement justifications

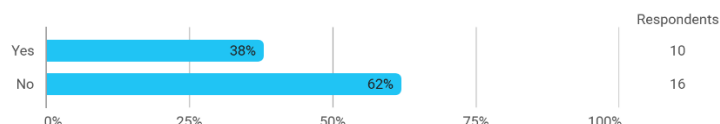
- This would change the test rational, outcome and usability
- More work and discussion is needed before the use of adapted inoculum is allowed
- pre-adapted inocula should not be used for these stringent tests

A.18 Test No. 302B: Inherent Biodegradability: Zahn-Wellens/ EVPA Test

Relevance of revision



Interest in being part of an OECD Expert Group

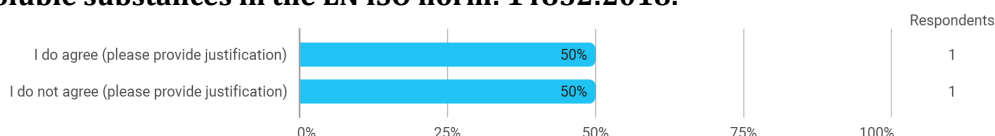


The following suggestions have been collected:

1. paras 2, 3 - Introduce a paragraph that allows the test to be conducted based on measurements of CO₂ formation

Rational: The OECD 302B test has been introduced for the testing of soluble substances under preferred conditions (high inoculum, high test concentration). Degradation is suggested to capture by DOC measurement.

The test was successfully amended and/or substituted by the measurement of CO₂, i.e. mineralisation (Strotmann et al., 1995). A comparison of results with outcomes from OECD 301F and OECD 302B led the authors to judge this test as giving reliable results and providing enhanced information on the pattern of biodegradation. This approach has been further developed for difficult substances by Gartiser et al. (2007). In fact, the principle of the test, including CO₂ measurement, has been adapted to insoluble substances in the EN ISO norm: 14852:2018.



Agreement justifications

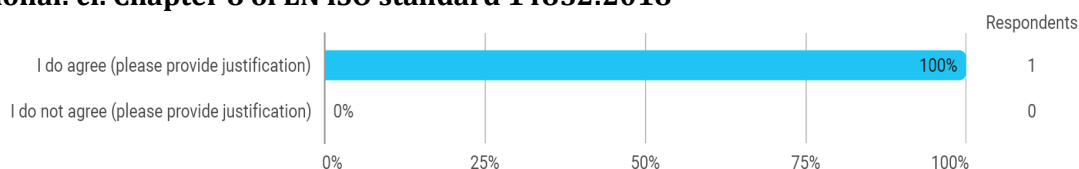
- Additional metrics for measuring degradation are needed for insoluble and/or adsorbing materials are needed

Disagreement justifications

- Is the measurement of CO₂ feasible given the high inoculum background signal?

2. para 5 - Adapt to non-soluble test substance and adopt techniques as described in EN ISO standard 14852:2018

Rational: cf. Chapter 8 of EN ISO standard 14852:2018

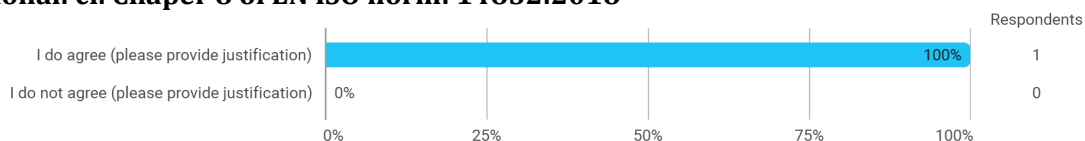


Agreement justifications

- Methods or adaptations are needed for handling poorly soluble and non soluble substances

3. paras 15-26: Adopt techniques as described in EN ISO norm: 14852:2018

Rational: cf. Chapter 8 of EN ISO norm: 14852:2018



Agreement justifications

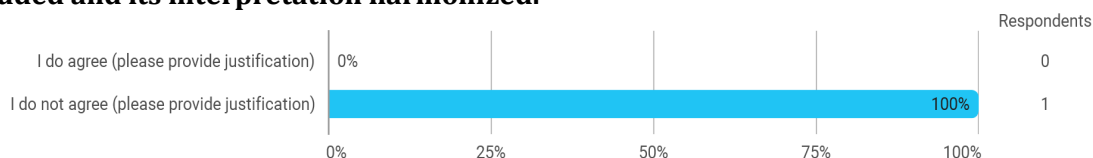
- Agreed, the methods need updated for handling polymeric materials

4. paras 28-30 - Add rationale for elimination and harmonize with REACH guidance document R.11 (ECHA, 2014)

Rational: ISO 9888 provides further guidance on the interpretation of the Zahn-Wellens test results: If the DOC?or COD-concentration after 3 hours is significantly lower than at the start (> 20% elimination), this is interpreted as abiotic elimination due to adsorption or volatilization.

The REACH guidance document R.11 (ECHA, 2017) is even stricter where the removed fraction before degradation occurs (usually 3-h value) must remain below 15% in a Zahn-Wellens test according to OECD 302B.

Overall the discrimination between elimination and biodegradation should be included and its interpretation harmonized.

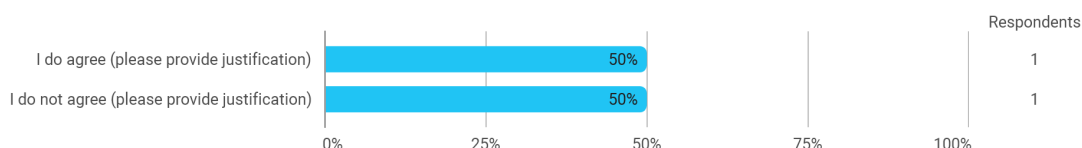


Disagreement justifications

- More clarification is needed on what in Chapter 11 they would like included, but I don't think the documents should cater to one geography

5. Additional reference compounds in both screening (301 series, 302) and simulations (307,308,309) tests should be included.

These additional reference compounds should allow to reflect degradability of more complex compounds ((bio)polymers).



Agreement justifications

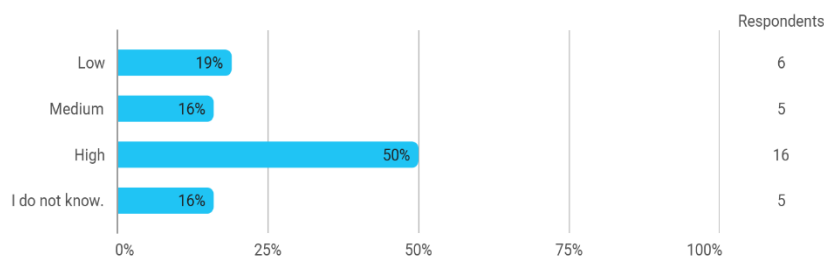
- More reference materials are needed, especially for polymers

Disagreement justifications

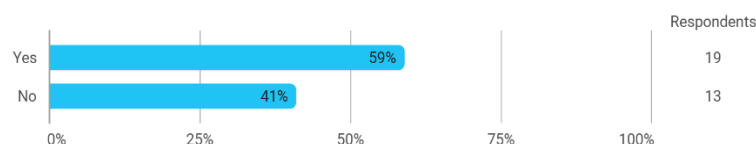
- Neither agree nor disagree: We don't understand this point.

A.19 Test No. 307: Aerobic and Anaerobic Transformation in Soil

Relevance of revision



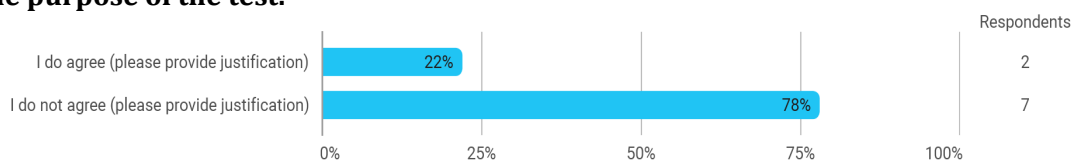
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. The test does not have any validity criteria and should include some. The test substance, if not degraded, should be proven to have stayed in the test medium throughout the test (10% loss at the end of the test allowed max). If not, the test should be considered invalid. Also, the test should also give some range of acceptable degradation kinetics of the reference item in function of the test temperature. Oxygen concentration should not fall below a certain level to be defined.

Rational: It is impossible to investigate degradability when the substance leaves the soil by evaporation. In that case, the study should not be considered valid because it is not fit for purpose. A more reliable and clear validity criterion is required for the degradation kinetics of the reference substance. The oxygen validity criterion is of utmost importance to ensure that the test accurately reflects an aerobic DT50, which is the purpose of the test.



Agreement justifications

- validity criteria and should include

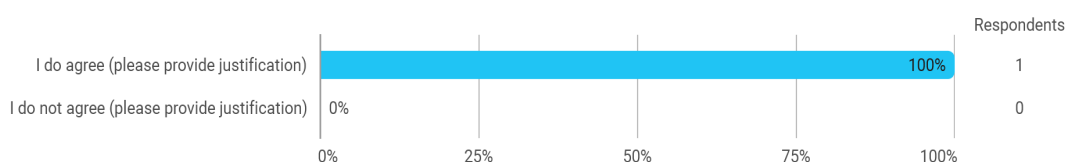
Disagreement justifications

- There are sufficient validity criteria already enshrined in the guidance, such as the requirement to achieve a mass balance of 90-110% for ¹⁴C studies and procedural recoveries of 70-110% for non-labelled substances. There is also a clear definition of acceptable microbial biomass for demonstrating soil viability
- This criteria becomes irrelevant in most cases where radiolabelled test substances are used, thus losses would be detected. In addition, according to current OECD 307 a validity criterion is already given, being an acceptable recovery range of 90 to 110% (radiolabelled test substances) and 70 to 110% (unlabelled test substances) of the applied amount. The

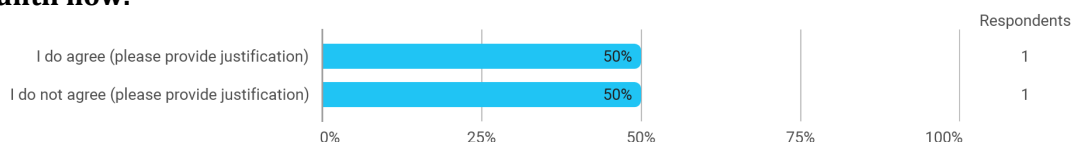
setup of the test ensures sufficient oxygen concentrations due to constant air flow in flow through systems and air exchange by diffusion in static test systems. Furthermore, the viability of the soil is tested by determination of the microbial activity.

- ▶ 307 contains already a lot of validity criteria; sufficient oxygen is already ensured by the test setup (e.g. flow trough)
- ▶ Aim of the study is to test 4 soils with different properties and activities!
- ▶ Do not think these restrictions are necessary. Suggest that reasonable limits of soil properties (E.g. organic carbon) are included.
- ▶ Context is not clear yet. According to current OECD 307 a validity criterion is already given, being an acceptable recovery range of 90 to 110% (radiolabelled test substances) and 70 to 110% (unlabelled test substances) of the applied amount. The setup of the test ensures sufficient oxygen concentrations due to constant air flow in flow through systems and air exchange by diffusion in static test systems. Furthermore, the viability of the soil is tested by determination of the microbial activity.
- ▶ Disagree, to my mind a validity criteria is already in place (acceptable recovery range of 90 to 110% for radiolabelled and 70 to 110 for non-radiolabelled substances). Reference compound not required and unsure what the reference to oxygen concentration relates to.

2. para 1 - Suggest adding in the introduction a paragraph explaining what the domain of applicability for this test guideline is (which types of substances it works best for)



3. para 1 - Suggest adding clear cut-off for Henry's constant, Koc, water solubility, etc, for which this test guideline does not apply. This could built on the properties of the substances on which this guideline was successfully vs unsuccessfully applied until now.

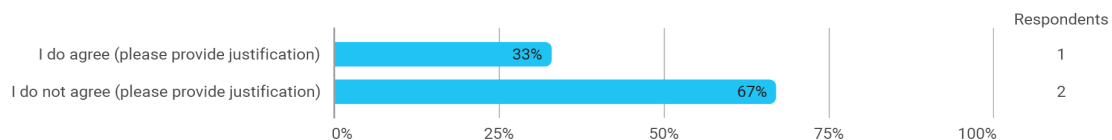


Disagreement justifications

- ▶ Some suggestions might be useful, but clear cut-off is not practicable because the combination of all properties is relevant to evaluate an appropriate test design

4. para 1 - It could be considered to add a procedural control to indicate the viability of the inoculum.

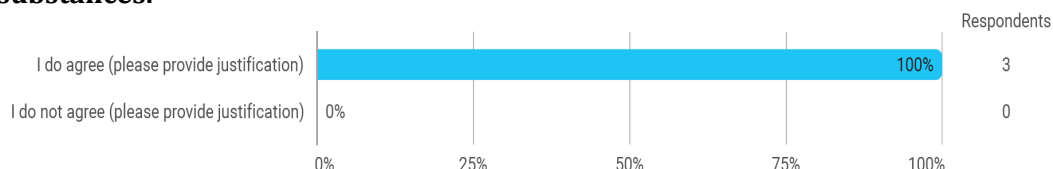
Rational: It is unclear why procedural control is not included in TG 307 although it is included in OECD TG 309. Viability of inoculum is important for all biodegradation tests and can be affected by sampling, storage, pre-treatment etc.



Disagreement justifications

- This is not necessary as there is a clear definition of an acceptable biomass using substrate induced respiration or fumigation extraction
- This is already in place!

5. para 5 - Suggest to make reference to the recent work of Shrestha et al. (2019) (<https://doi.org/10.1021/acs.est.8b05079>) concerning the testing of volatile substances.

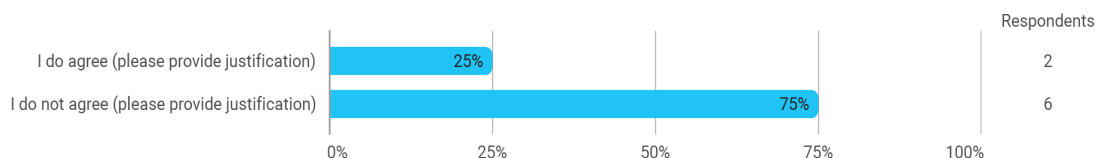


Agreement justifications

- with some kind of advice
- Adding an internal trap for volatiles must be evaluated very carefully, the trap can act as kind of "magnet"
- Testing of slightly volatile substance.

6. paras 12-14 - Extraction Efficiency shall ensure that method and solvent used is adequate for its purpose. this is the case if the extraction method is capable to retrieve at least 90 % of the substance in the first extraction step and less than 5 % remains in the last extraction step.

Rational: I propose to add a para dealing with efficiency of the method used. Extraction efficiency directly influences reliability of data. Only if the extraction method is efficient enough it makes sense to review the recovery or, with other words, it is a precondition if the recovery rate shall make sense. It should thus be defined and reported.



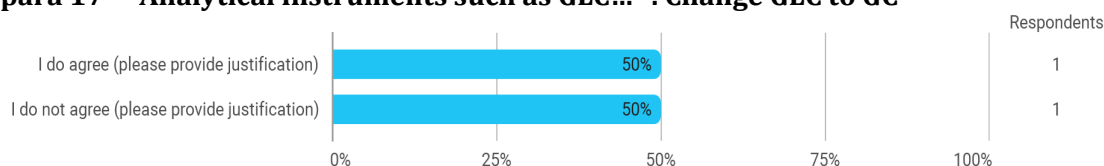
Agreement justifications

- disagree for 90% in the first step but agree if 90% in total is acceptable after the last step
- Extraction methods like ASE/PLE should be mandatory

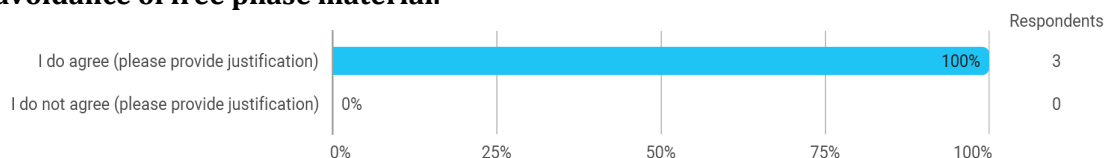
Disagreement justifications

- ▶ Extraction of >90% in the first extraction step is rarely achievable. The key point is that there should be > 90% extractability (from the entire extraction scheme) in the zero-day samples to demonstrate that this is a suitable method for use for the rest of the sampling points in the study
- ▶ The current aim of study conduct is to extract >90% at 0 Days with all extraction steps to define the suitable extraction regime for the rest of the incubation. As degradation progresses, we often run into instances where the first extraction step does not pull out 90% AR. This criteria would be extremely difficult to meet and is a bit arbitrary. As long as the mass is recovered and profiled, why does the distribution across the extraction steps matter?
- ▶ stability of the test compound in the extraction solvents and sufficient extractability needs to be shown anyway at the end of the extraction procedure incl. extraction at higher temperature if applicable. Introducing such triggers per extraction step is unrealistic and does unnecessarily complicate the lab work and may (dependent on the substance properties) not be feasible.
- ▶ It is unclear what would be appropriate if NERs are more than 10%
- ▶ The extraction efficiency and stability of the test item in extracts has already to be shown at the end of the extraction procedure. Therefore, an introduction would only unnecessarily increase complexity of a study. The aim should be to have an extraction efficiency at Day 0 of >90% for the whole extraction procedure.
- ▶ Strongly disagree on this point. The amount extracted in the first extract is irrelevant to the suitability of the method. Quick extraction is desirable practically, but a larger number of lower efficiency steps should be acceptable as long as total extractability is acceptability and stability maintained throughout the longer extraction process.

7. para 17 - "Analytical instruments such as GLC... ". Change GLC to GC



8. para 20 - It should be noted that different dosing methods can have a significant impact on the bioavailability of the test substance, and hence on the observed degradation. For example, poorly soluble solid compounds added as such are likely to have significantly reduced bioavailability (due to reduced surface area) compared with those added with a solvent. Wherever possible it should be sought to have homogeneous distribution of the test substance in the system and the avoidance of free phase material.

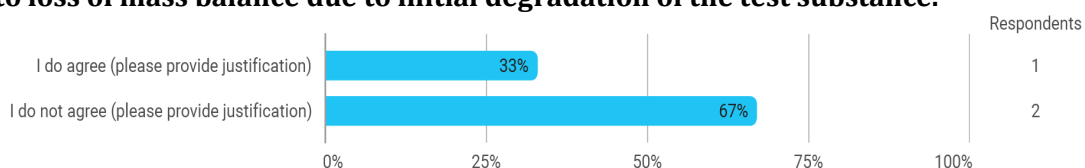


Agreement justifications

- ▶ However "Wherever possible" should be more precise as well as "homogeneous"

- Homogeneous distribution is already required by the guideline

9. **para 21 - Suggest removing this paragraph- this is not a common practice in standard regulatory testing. The direct spiking of sludge does not effectively simulate in situ degradation in sludge, where typically chemicals are present as a residue with significantly reduced bioavailability. Spiking of sludge may also lead to loss of mass balance due to initial degradation of the test substance.**



Agreement justifications

- this is not a common practice

Disagreement justifications

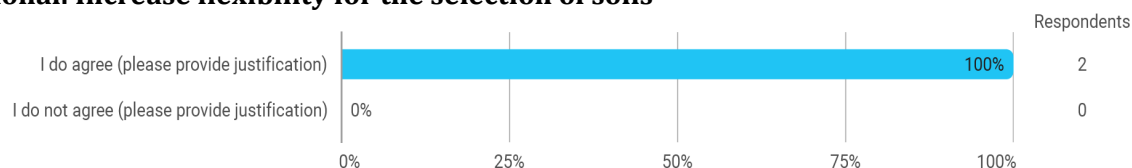
- this is directly applicable for chemicals entering the soil compartment after wastewater treatment (on biosolids). If this is a dominant route of exposure for a specific chemical then dosing on sludge solids is environmentally relevant and should be done in this type of simulation study.
- For veterinary drugs this method was requested in the past

10. **paras 3-24 - Suggest revising these paragraphs to make more clear the specifications required for soils under the different test variants (transformation rate vs pathway studies). Experience has shown that the initially required soil characteristics can cause confusion for the more common transformation rate studies, where more flexibility in soil characteristics is afforded.**



11. **para 26 - The statement "If soils have been treated with the test substance or its structural analogues within the previous four years, these should not be used for transformation studies (10)(12)" is too restrictive with regard to the possibilities to find representative arable soils under agricultural use. The metabolism and speed of degradation of chemicals is only in exceptional cases affected by a treatment of similar compounds in previous years. And these are usually well known chemical classes like phenox acids. It would facilitate the selection of soils considerably if the 4-years rule is relaxed to e.g. 1-2 years.**

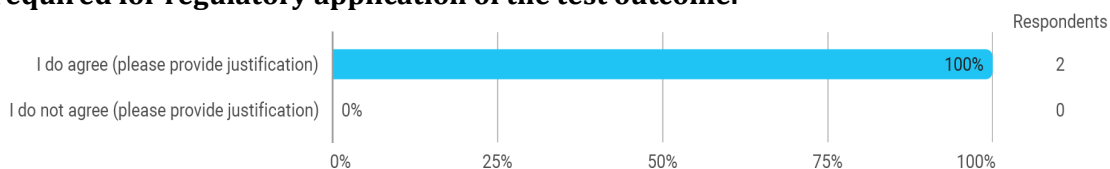
Rational: Increase flexibility for the selection of soils



Agreement justifications

- application changes the microbial community with long lasting effects; there should be no shortening of this period.

12. para 31 - Add in this section "There may be different incubation temperatures required for regulatory application of the test outcome."

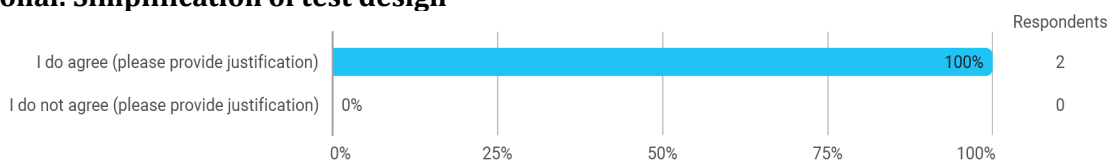


Agreement justifications

- ▶ different incubation temperatures required

13. para 32 - The request "For chemicals applied or released in colder climates (e.g. in northern countries, during autumn/winter periods), additional soil samples should be incubated but at a lower temperature (e.g. 10 ± 2 °C)" is superfluous and can be deleted as the soil metabolism is hardly affected by lower temperatures. It is acknowledged that the speed of degradation will be affected but that can be accounted for by agreed recalculation procedures

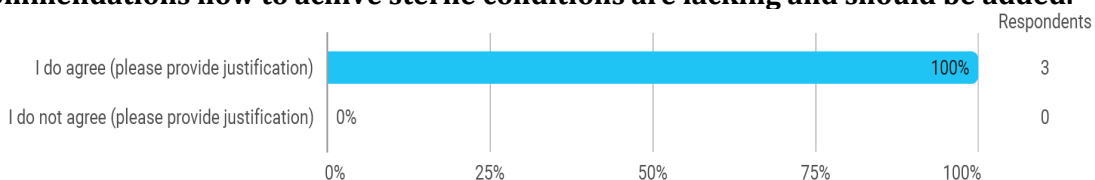
Rational: Simplification of test design



Agreement justifications

- ▶ para can be removed
- ▶ Fully agree. This recommendation is irrelevant. It is up to appropriate regulations to define such a requirement

14. para 36 - sterile controls deliver important, sometimes decisive information
Rational: Sterile controls should not be optional but mandatory. They produce important information that often are decisive for interpretation especially if the test substance is difficult to test, e.g. very low water solubility or volatility. Clear recommendations how to achieve sterile conditions are lacking and should be added.

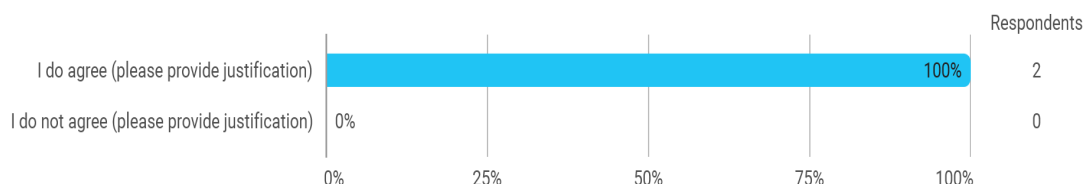


Agreement justifications

- ▶ only for non-labelled substances
- ▶ We agree, but defining appropriate procedure to obtain a good sterile control is not that straightforward and may deserve further research work
- ▶ Agree that sterile controls can deliver important information,

15. para 36 - The irrelevant references (13, 26) indicated for soil sterilisation methods should be deleted and replaced with useful references.

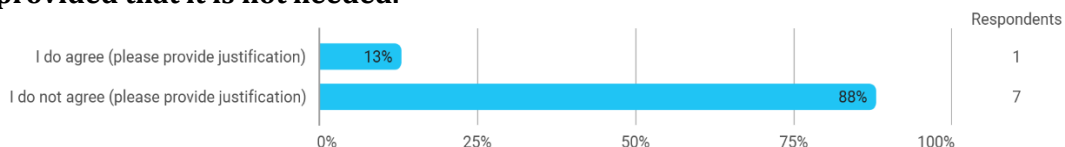
Rational: The references 13 and 26 are irrelevant for the indicated purpose (soil sterilisation methods). They do not include any advice on soil sterilisation methods.



Agreement justifications

- We potentially agree, but appropriate references have not been proposed

16. para 36 - The sterile controls should be made mandatory, unless a justification is provided that it is not needed.



Agreement justifications

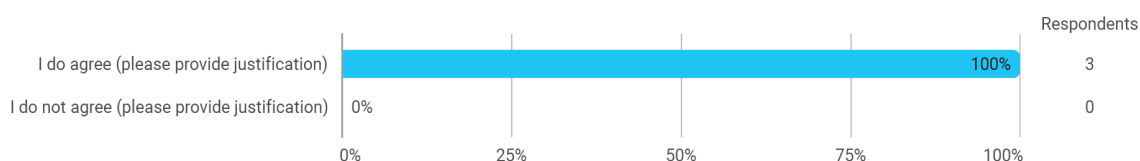
- more details are needed (e.g. sterilisation method, sampling points, replicates)

Disagreement justifications

- Sterile controls can be useful but only under specific circumstances (such as supporting the formation of biogenic NER). However, in many cases there would be little value in conducting the tests and inclusion would simply add additional complexity and cost for no benefit. Therefore, it is more appropriate to keep this as an optional test to be used on a case-by-case basis
- Testing in sterile systems should remain optional. The requirement of a sterile system would be a discussion for registration data requirements and not the OECD guideline itself
- sterile control may be useful for explaining the degradation mechanism; they do not help to derive degradation rate, therefore they should be left optional
- Sterile controls should be left optional.
- There is a debate to identify relevant methods for sterilization. This debate illustrates that the sterilization procedure is not straightforward. That should be left optional for the moment.
- The inclusion of sterile controls should stay optional and inclusion should be decided case-by-case.
- Disagree that sterile controls should be compulsory in the guideline (that is a matter for national regulatory data requirements)

17. para 37 - For definition of anaerobicity, usually the redox potential value of -100 mV (Eh) is requested by evaluators although not explicitly mentioned in this guideline. This value is taken from the OECD 308 guideline which deals with sediment, where those values may occur. For soil incubations, this value is considered unrealistic. The necessity for measuring redox potential should therefore be deleted.

Rational: Rejection of studies/ experiments should be avoided in future just because this unrealistic value (cross-referenced from another guideline) is applied.

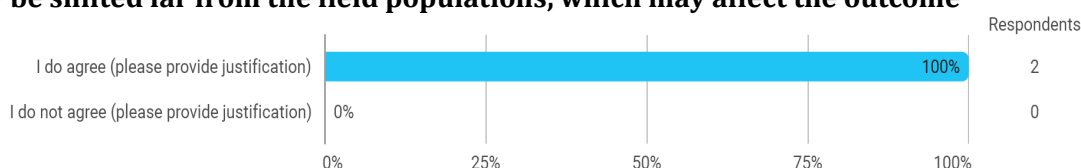


Agreement justifications

- For soil incubations, this value is considered unrealistic. Is there a more useful parameter? If not, only the procedure of nitrogen addition should be sufficient (without checking any type of parameter)
- In an agricultural context, it is unlikely that field soils would extremely be anaerobic for an extended period of time and it would be extremely difficult to establish a suitable redox value that represents realistic anaerobic conditions. As such, presentation of such data adds little value to the study, if it cannot be contextualized
- Agree, maintenance of the units in an anaerobic environment should be sufficient (and this is relatable to the true situation in the field where soils subject to flooding will not immediately become anaerobic).

18. para 37 - The test conditions for the anaerobic incubation are extreme relative to field conditions (120 days continuous anaerobicity rather than transient periods of anaerobicity as indicated in the footnote on page 6). A design that is more representative of field conditions would be of more value.

Rational: There is a contradiction between the extreme conditions in the OECD design and field conditions which makes it difficult to contextualise the outcomes in terms of significance to the field (DT50 and metabolite levels). Furthermore, by forcing the system to an extremely low redox potential the population of anaerobic organisms may be shifted far from the field populations, which may affect the outcome

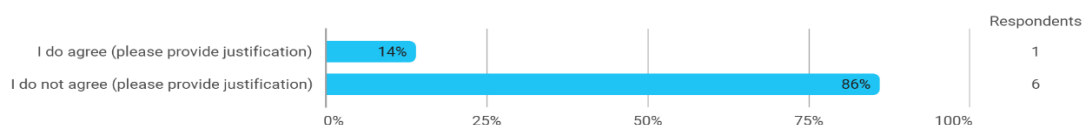


Agreement justifications

- Anaerobic conditions in agricultural soils are likely to be transient in nature (following heavy rain) and, therefore, incubation of the system for 120 days completely under extreme anaerobic conditions does not represent field conditions. Compounds are likely to be exposed to more intermediate redox conditions and short periods of time and cycle between aerobic and anaerobic conditions. The current test design is likely to lead to a significant misrepresentation of the significance of anaerobic degradation
- Agree a more realistic anaerobic study design may be desirable, but unsure what this would look like.

19. para 39 - The rate and pathway studies should last more than 120 days. At best 180 days.

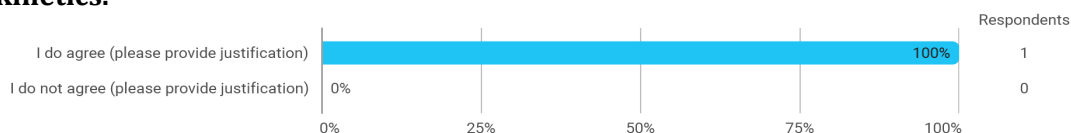
Rational: In persistence assessment 120 days is the trigger for persistence and 180 Days the trigger for very persistent substances. Experience shows that the system usually remains stable for 120 days and beyond. The Shorter the time frame is to 180 days the less Extrapolation is needed or even a direct comparison may be possible. With smaller Extrapolation conclusion becomes more certain.



Agreement justifications

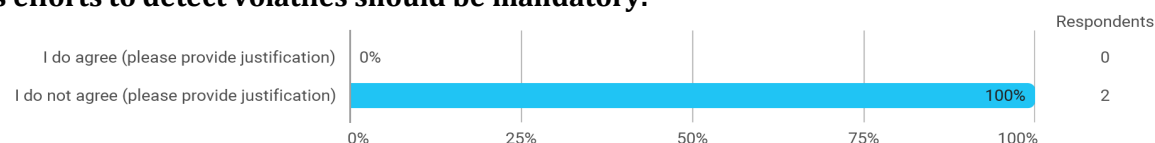
- ▶ Soil viability must be evaluated carefully if the study is running > 120 days
- ▶ Suggestion 19 - disagreement justifications
- ▶ Maintaining soil viability for up to 120 days can already be challenging (there is often a significant decline in biomass over this time frame). Extending the duration would simply exacerbate this issue
- ▶ There is significant difficulty in maintaining soil viability for longer incubation periods. Regulatory reviewers frequently reject data collected beyond 120 days due to soil viability concerns
- ▶ there is good reason not extend study beyond 120 d (microbial viability); prolongation can be optional, but results must then be considered with care
- ▶ We don't see the need to go for at least 120 days if the relevant information is obtained after 60 days. No opinion on the possibility to extend to 180 days. It is our understanding that the viability of the inoculum is not guaranteed. This may be temperature dependent
- ▶ Regulators often reject data points in studies conducted for more than 120 days due to soil viability issues, thus this proposal should not be considered.
- ▶ Disagree, incubation beyond 120 days compromises soil viability and can result in erroneous conclusions of biphasic behaviour (arguably this is already the case at 120 days).

20. para 41 - Suggest modification of this paragraph to recommend for persistence assessment that substances are dosed at a sufficiently low concentrations to reduce the presence of free phase material, and derive appropriate degradation kinetics.



21. para 45 - Sampling Points are too few in case of difficult to test substances; Detection of volatile substances (parent and metabolites) should be a mandatory request.

Rational: In case of quickly dissipating substances with slow Degradation for assessment it is necessary that sufficient data Points are available
Overall recovery strongly depends on volatile substances, impacts the validity and thus efforts to detect volatiles should be mandatory.



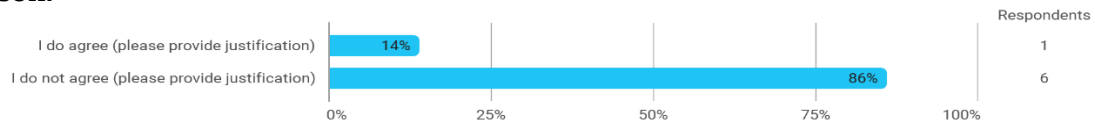
Disagreement justifications

- ▶ There is no restriction for number of sampling points, detection of volatiles is already in the guideline and necessary for mass balance calculations

- ▶ Neither agree nor disagree: §45 does not deal with the number of sampling points

22. para 46 - Extraction of soil with a proper solvent should be repeated until the extract contains less than 5% of the first extract.

Rational: The 5% measure endures that most of the extractables are removed from the soil.



Agreement justifications

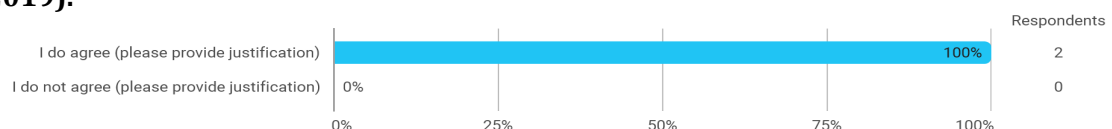
- ▶ See suggestion 6

Disagreement justifications

- ▶ The wording on the current guideline is sufficient. The choice of extraction solvent will be compound-specific and, as mentioned previously, the most important point is the > 90% is extracted at 0-day with the entire extraction method not with the single first extraction step (which will rarely be achievable)
- ▶ this concerns the quality of the analytical method for which own guidelines exist
- ▶ Unclear statement.
- ▶ Reasoning in percentage can be tricky. If the first extraction yields 1%, it is questionable whether a second extraction should be performed
- ▶ Context is not clear. What is meant by “proper solvent”?
- ▶ Unsure what is meant by "proper solvent". Generally extraction would be performed until the extract contains less than 5% applied not 5% of the first extract.

23. para 46 - Update to reflect newer methods for the determination of the biomass in the control test vessels

Rational: The test guideline is standardized in many ways, but least so with respect to determination of microbial biomass, which is the catalyst of the system. Many of the current methods employed are > 70 years old. There are accurate and precise newer methods routinely used in microbiology, which get closer to measuring the catalytic element of biodegradation in the microbial biomass, enabling standardization/normalization of the determination or providing further information with which to interpret outcomes. For example quantitative real-time polymerase chain reaction (PCR) targeting universal marker genes (Harms et al. 2003), total cell counts using epifluorescence microscopy (EFM) or flow cytometry (FCM). Total cell counts using fluorescent staining of DNA with EFM is traditionally considered the “gold standard” for quantifying bacterial cells in the aqueous environment (Brown et al. 2019).

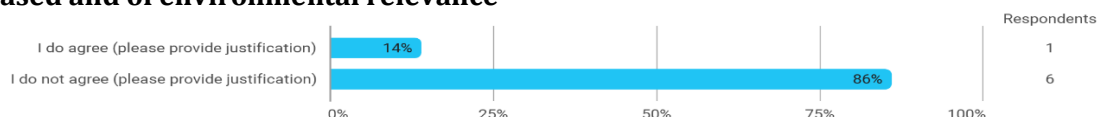


Agreement justifications

- Existing methods (SIR, fumigation extraction) are old and whilst they generate values that indicate overall viability it is not possible to link decline in biomass to decline in degradation rate (cause and effect)) over time. Maybe there are other optional methods that would provide more information

24. para 47 - Beside NER quantification, NER should be characterized with respect to the three different types: sequestered, covalently bound, biogenic. See Schäffer et al. (2018), ESEU 30:51

Rational: Information of sequestered residues is important: type 1 NER is slowly released and of environmental relevance



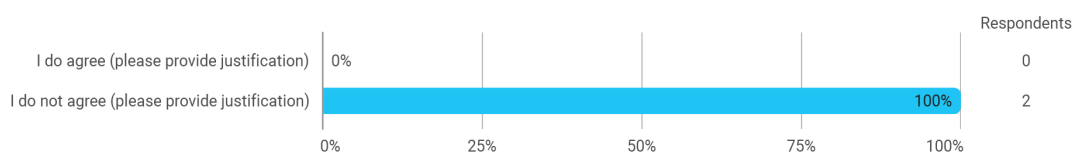
Agreement justifications

- NER need to be updated. However with an applicable way. Silylation could be difficult to perform in a lab

Disagreement justifications

- Different regional regulatory authorities have their own interpretation of NER and their own proposals on how to assess their significance (e.g. the US EPA extraction efficiency guidance, 2014). The proposed methodology addresses a specific EU interpretation of NER and, as such, it would not be appropriate to include this in an international guidance. The ECHA will ultimately issue separate guidance, or it can be included in specific data requirements legislation as required. Furthermore, the methods proposed are not fully validated or broadly available in many laboratories, therefore, the time is not right to consider this step
- This requirement would over complicate the conduct of the study for many regions outside of EU that rely on OECD guidelines. Also, the technical feasibility of the Schaffer et al. proposal has not been fully realized yet. This proposal is still under stakeholder review
- this is an analytical procedure and has nothing to do with harmonizing the test setup for which the guideline is aiming for; the need for those investigations should be optional and left to the data requirements of the legislation under which the study has to be performed
- Strongly disagree. The science is not ready for this. A ring-test is needed before inclusion of such a heavy protocol in a TG.
- Although scientifically reasonable, the need for NER characterization should be optional and left to the data requirements of the legislation under which the study has to be performed. The current approach of NER characterization is rather EU centered.
- NER characterisation should be a matter for national data requirements (for example the US EPA guidance on the subject)

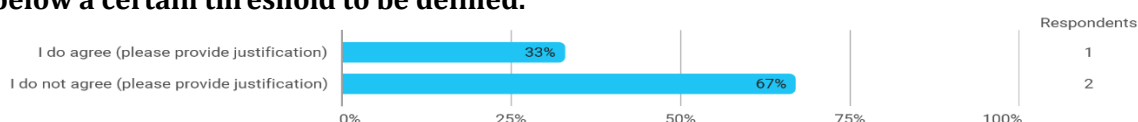
25. Additional reference compounds in both screening (301 series, 302) and simulations (307,308,309) tests should be included. These additional reference compounds should allow to reflect degradability of more complex compounds ((bio)polymers).



Disagreement justifications

- This is quite difficult
- Reference compound not required for the 307 study

26. validity data - The test should include validity criteria. Quality criteria are not enough because they are used to assess the validity of a study. One criterion could be that the test substance has stayed in the sediment/water system. If not, the test is not valid. Also, the test should also give some range of acceptable degradation kinetics of the reference in function of the test temperature. Oxygen should not fall below a certain threshold to be defined.



Agreement justifications

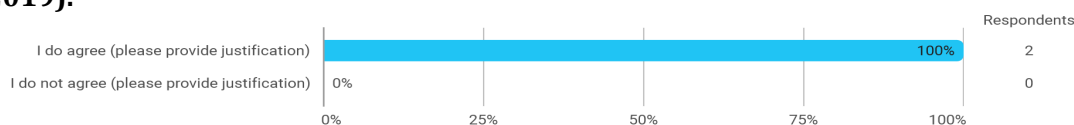
- There is no validity criteria in this TG. That is a strong issue that prevents unreliable studies to be repeated.

Disagreement justifications

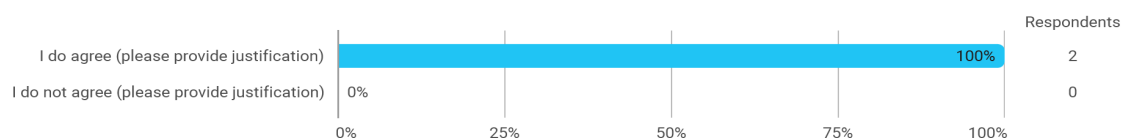
- Seems to be a copy/paste error. Refers to wate-sediment systems
- Repetition of Suggestion 1 - see above

27. Update to reflect newer methods for the determination of the biomass in the control test vessels

Rational: The test guideline is standardized in many ways, but least so with respect to determination of microbial biomass, which is the catalyst of the system. Many of the current methods employed are > 70 years old. There are accurate and precise newer methods routinely used in microbiology, which get closer to measuring the catalytic element of biodegradation in the microbial biomass, enabling standardization/normalization of the determination or providing further information with which to interpret outcomes. For example quantitative real-time polymerase chain reaction (PCR) targeting universal marker genes (Harms et al. 2003), total cell counts using epifluorescence microscopy (EFM) or flow cytometry (FCM). Total cell counts using fluorescent staining of DNA with EFM is traditionally considered the “gold standard” for quantifying bacterial cells in the aqueous environment (Brown et al. 2019).



28. Annex 4 - Add illustration of test system described in Shrestha et al 2019 for testing of volatile compounds.

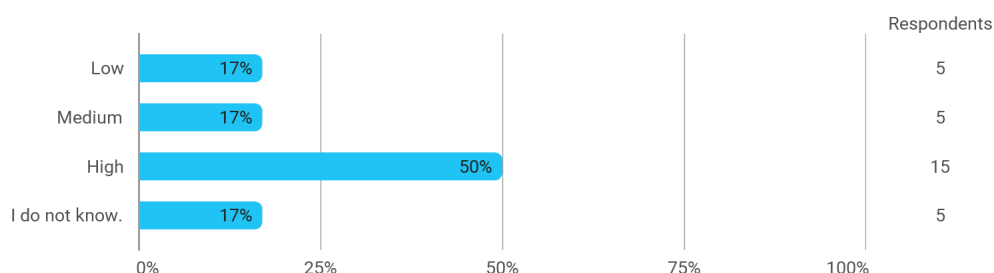


Agreement justifications

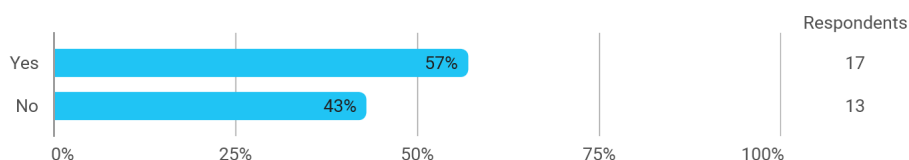
- But this is only one example, the optimised test design must be evaluated for each substance individually

A.20 Test No. 308: Aerobic and Anaerobic Transformation in Aquatic Sediment Systems

Relevance of revision



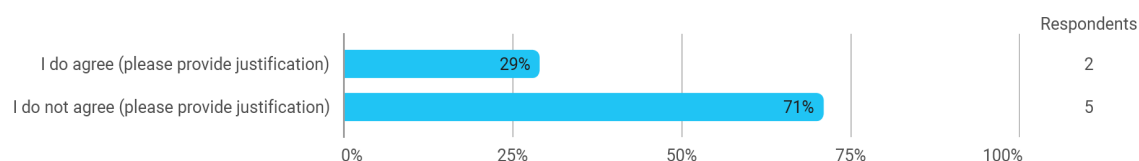
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

- The test does not have any validity criteria and should include some. The test substance, if not degraded, should be proven to have stayed in the test medium throughout the test (10% loss at the end of the test allowed max). If not, the test should be considered invalid. Also, the test should also give some range of acceptable degradation kinetics of the reference item in function of the test temperature. Oxygen concentration should not fall below 7 mg/L in the water phase.**

Rational: It is impossible to investigate degradability when the substance leaves the sediment/water system by evaporation. In that case, the study should not be considered valid because it is not fit for purpose. A more reliable and clear validity criterion is required for the degradation kinetics of the reference substance. The oxygen validity criterion is of utmost importance to ensure that the test accurately reflects an aerobic DT50, which is the purpose of the test.



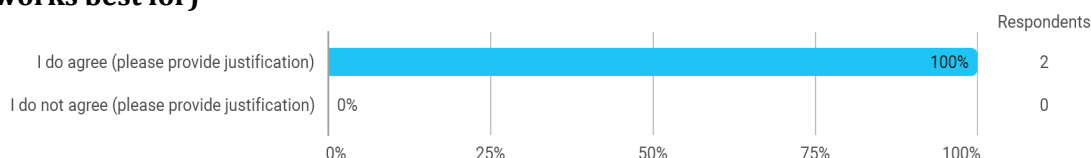
Agreement justifications

- Fully agree.

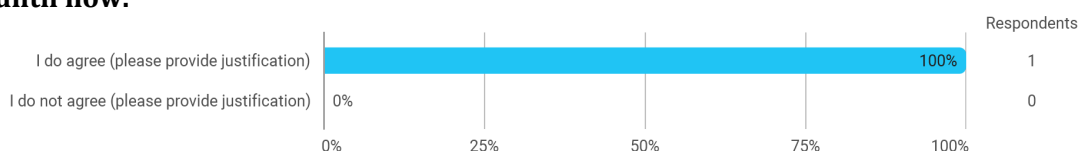
Disagreement justifications

- ▶ Sufficient validity criteria already exist, such as a mass balance requirement of 90-110% for ¹⁴C studies and procedural recoveries of 70-110% for non-labelled compounds. In addition an active biomass should be demonstrated
- ▶ In most cases radiolabeled test substances are used, thus losses would be detected. In addition, according to current OECD 308 a validity criterion is already given, being an acceptable recovery range of 90 to 110% (radiolabeled test substances) and 70 to 110% (unlabeled test substances) of the applied amount. Furthermore, the viability of the sediment is tested by determination of the microbial activity.
- ▶ there are validity criteria provided (recoveries of radioactivity or test item, microbiological activity etc.)
- ▶ According to current OECD 308 a validity criterion is already given, being an acceptable recovery range of 90 to 110% (radiolabeled test substances) and 70 to 110% (unlabeled test substances) of the applied amount. Furthermore, the viability of the sediment is tested by determination of the microbial activity.
- ▶ Validity criteria already established and well understood (range 90 - 110% for radiolabelled test substances and range 70 - 110% for non-radiolabelled test substances). Reference item (in the meaning here) not required for the 308 study therefore its kinetics not relevant)

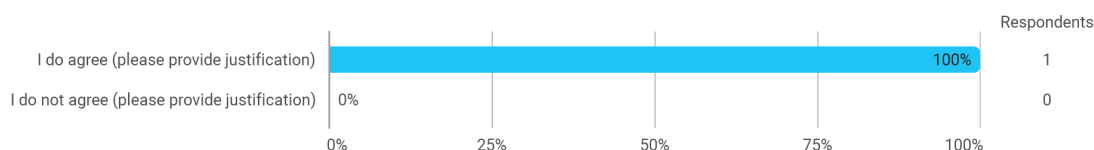
2. para 1 - Suggest to add in the introduction a paragraph explaining what the domain of applicability for this test guideline is (which types of substances it works best for)



3. para 1 - Suggest to add clear cut-off for Henry's constant, Koc, water solubility, etc, for which this test guideline does not apply. This could built on the properties of the substances on which this guideline was successfully vs unsuccessfully applied until now.

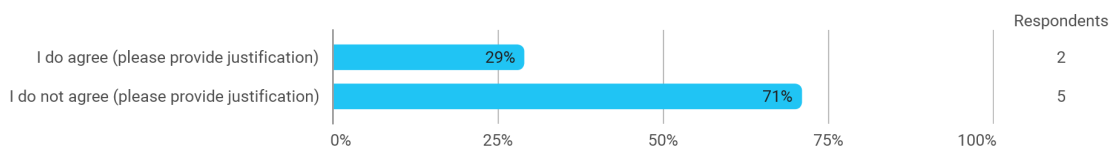


4. para 5 - Additional sentence: Chemicals that are both hydrophobic and volatile are also likely to present challenges and should be considered with care when testing according to this guideline.



5. para 2 - Recommendation/requirement for sterile control bottles should be added, with sufficient technical advice e.g. for sterilisation of the water-sediment systems. It could be considered to add a procedural control to indicate the viability of the inoculum.

Rational: This test is used e.g. under REACH regulation to determine degradation/transformation half-lives. There is a need to differentiate degradation from other loss processes. Sterile controls are needed for this. It is unclear why procedural control is not included in TG 308 although it is included in OECD TG 309. Viability of inoculum is important for all biodegradation tests and can be affected by sampling, storage, pre-treatment etc.



Agreement justifications

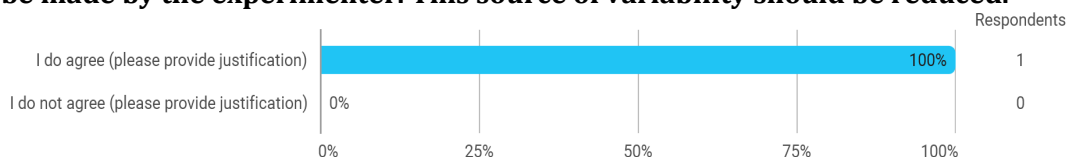
- Sterilisation process need to be stated in more detail
- Fully agree.

Disagreement justifications

- Sterile controls should only be considered an optional test as, in most cases, they will not add sufficient value to justify the added complexity and cost of the study
- Testing in sterile systems should remain optional. The requirement of a sterile system would be a discussion for registration data requirements and not the OECD guideline itself
- sterile controls should be left optional; they produce sometimes strange results which are not representative for environmental conditions
- Should be included as an optional test only, nevertheless clear advice on sterilisation procedures should be provided nevertheless.
- Should not be a requirement, but included as an optional test.

6. para 18 - The dimensions of the apparatus must be such that the requirements of the test are complied with (see paragraph 32) and should have a base area of x (to be defined) cm² and height to accommodate the water volume needed to accommodate what is defined in paragraph 32.

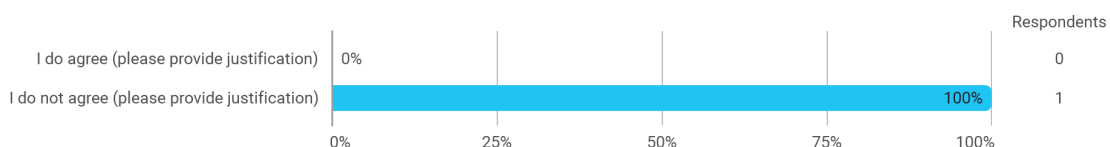
Rational: It was reported at the final workshop of the Cefic funded project LRI ECO18 improved strategy to assess chemical persistence at the water sediment interface 'Identifying limitations of the OECD water sediment test (OECD 308) and developing alternatives to assess persistence' held in Duebendorf Switzerland in October 2015, that one factor that has a large impact on the results obtained in the study is the geometry (height to base dimensions) of the test vessel. Therefore it seems necessary to define this more clearly to improve confidence in and reduce variability in endpoints produced from the study that originates from a parameter that is not related to variability in a true environmental parameter but to a choice that can be made by the experimenter. This source of variability should be reduced.



Agreement justifications

- Standardization is needed.

- 7. para 18 - Examples of typical test apparatus, i.e. gas flow-through and biometer-type systems, are shown in Annexes 3 and 4, respectively (11). Other useful incubation systems are described in reference 12. The design of the experimental apparatus should permit the exchange of air or nitrogen and the trapping of volatile products. The dimensions of the apparatus must be such that the requirements of the test are complied with (see paragraph 32). Ventilation may be provided by either gentle bubbling or by passing air or nitrogen over the water surface. In the latter case gentle stirring of the water from above may be advisable for better distribution of the oxygen or nitrogen in the water. CO₂-free air should not be used as this can result in increases in the pH of the water. In either case, disturbance of the sediment is undesirable and should be avoided as far as possible. Slightly volatile chemicals should be tested in a biometer-type system with gentle agitation of the water surface by either overhead stirring or by placing the sample on an orbital shaker (set to speed around 80rpm). Closed vessels with a headspace of either atmospheric air or nitrogen and internal vials for the trapping of volatile products can also be used (13). Regular exchange of the headspace gas is required in the aerobic test in order to compensate for the oxygen consumption by the biomass**

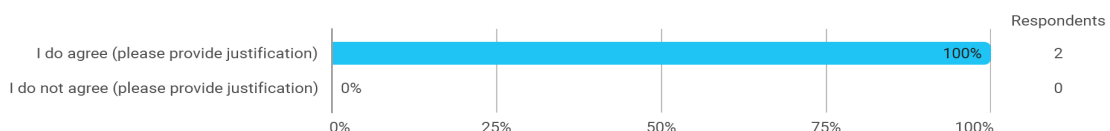


Disagreement justifications

- Partly disagree. From our experience, it is clear that magnetic stirring doesn't work.

- 8. paras 13-15 - Add a request to measure and report extraction efficiency and give a definition.**

Rational: Extraction efficiency directly influences reliability of data. Only if the extraction method is efficient enough it makes sense to review the recovery or, with other words, it is a precondition if the recovery rate shall make sense. It should thus be defined and reported. This shall ensure that method and solvent used is adequate for its purpose.



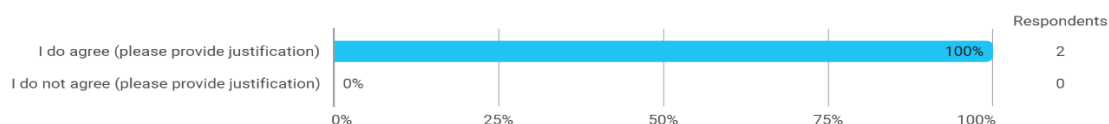
Agreement justifications

- definition needed

- 9. paras 17-18 - Test vessel geometry is reported to influence test results and should thus be exactly defined**

Rational: Test vessel geometry is reported to heavily influence results. Test results should not depend on the test vessels used else it is impossible to reliably conclude on degradation properties. This has grave consequences for the usefulness of the test

system that is routinely used to assess substances in the context of different legislations

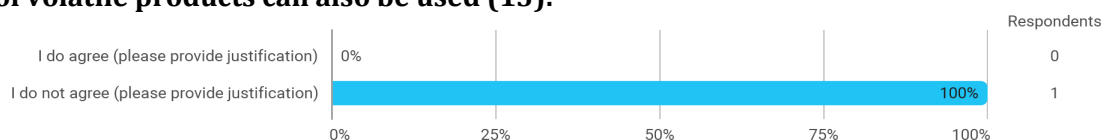


Agreement justifications

- definition needed
- There is indeed a need for a better standardization. However, the parameters should be scientifically justified, as these are key. Unclear whether we have the science to support this or that geometry and size.

10. para 18 - It should be made explicitly clear that agitation of the water phase is required in all closed systems to facilitate the exchange of oxygen between headspace and water.

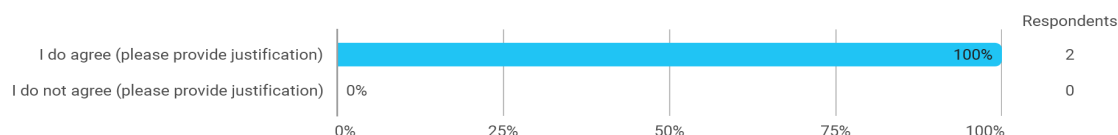
Revise the sentences: "Slightly volatile chemicals should be tested in a biometer-type system with gentle stirring of the water surface. Closed vessels with a headspace of either atmospheric air or nitrogen and internal vials for the trapping of volatile products can also be used (13)."



Disagreement justifications

- We agree, but the suggestion to use stirring is not relevant. Stirring does not work. Only shaking may do so. It is of paramount importance to monitor O₂ levels in the water phase, to check it does not fall below 7 mg/L (aerobic test).

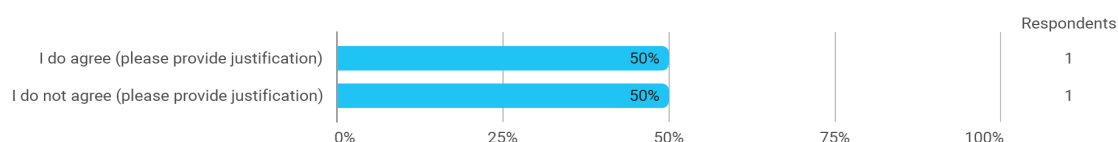
11. para 18 - Include the sentence: "Instead of stirring the water surface, exchange of air between headspace and water phase can also be facilitated by gentle shaking of the test systems. Shaking can be conducted using orbital shakers, and must be at a minimum (< 80 rpm) to prevent disturbance of the sediment layer. This can be assessed by measuring turbidity of solutions close to the sediment layer."



Agreement justifications

- How should turbidity be measured?
- No specific comment

12. para 19 - Include Tenax as a suitable absorbent for trapping volatiles.



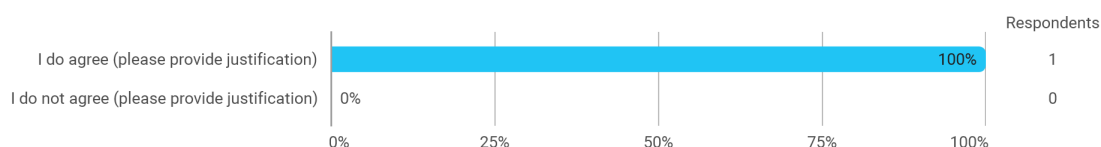
Agreement justifications

- We agree if supported by scientific findings.

Disagreement justifications

- Or stay as unspecific as possible for further trapping content. The trapping material could differ from test to test.

13. para 27 - Suggest a recommendation to include non-invasive optical oxygen sensors to measure oxygen concentrations in both headspace and water.

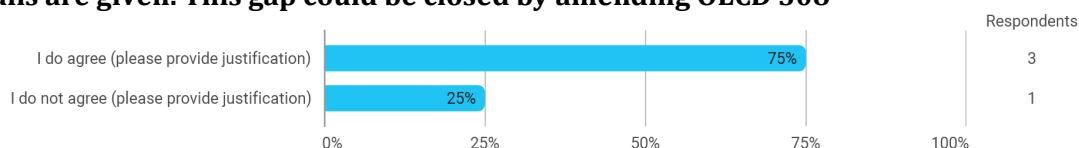


Agreement justifications

- Fully agree. O2 should be validity criterion for this TG

14. para 33 - Include as option the irradiated water-sediment study ; Include (in addition to the study under dark conditions) the higher-tier option to run an irradiated water-sediment study and provide some key experimental conditions (e.g. specification that aerobic conditions only are required, use of a Xenon arc lamp and quartz vessels with same requirements as given in OECD 316); these experimental details could also be given in an additional Appendix.

Rational: An irradiated water-sediment study can provide useful additional information on the environmental behaviour of compounds for which photochemical degradation is of importance. Such a study is explicitly mentioned as a higher-tier option in Commission Regulation (EU) No. 284/2013, Point 9.2.3, but no experimental details are given. This gap could be closed by amending OECD 308



Agreement justifications

- Agree with specific details how to perform.
- The irradiated water-sediment study (under UV light) is an optional study under the EU data requirements and can provide valuable information of the impact of photodegradation in a more realistic system as part of a weight of evidence assessment of persistence. This would be a good opportunity to incorporate this into the OECD308 guideline as an optional test
- Agree, the addition of specifications for an irradiated water sediment would be useful.

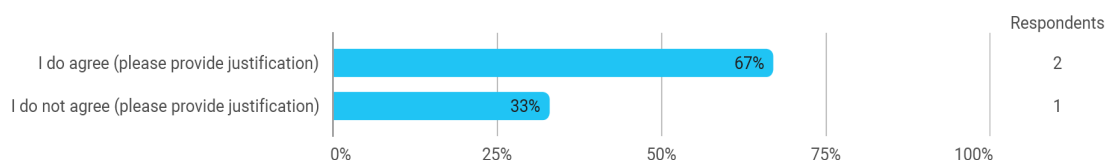
Disagreement justifications

- Neither agree nor disagree: The added value of such an option is unclear in the suggestion.

15. para 33 - There is significant research in the literature on the metabolic competency of algae and aquatic macrophytes. The potential contribution of these organisms is excluded when studies are conducted in the dark. The inclusion of a higher tier option to conduct a water-sediment study under the same lighting

regime as used for algal ecotoxicology studies (i.e. fluorescent PAR light) would be beneficial to the understanding of the behaviour of the compound. This would be analogous to the use of “diffuse light” in OECD 309 studies. The use of fluorescent light would exclude UV wavelengths and allow differentiation between this test and the irradiated water sediment study which is also permitted in the EU data requirements.

Rational: The current test design excludes a critical part of the microbial biomass, both in terms of direct metabolic competency, but also in terms of indirect effects that the presence of phototrophic organisms has on the structure of the heterotrophic microbial community. Hence, the true persistence of the test compound could be misrepresented by a study performed in the dark.



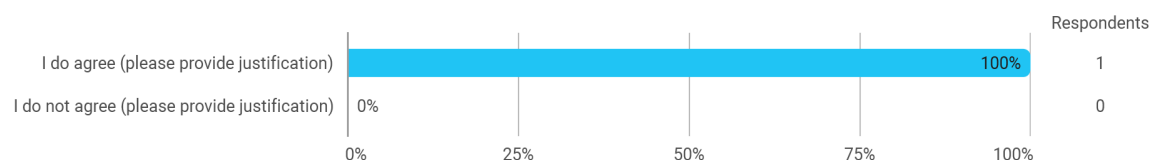
Agreement justifications

- Incubation in continuous darkness prevents both direct xenobiotic metabolism by phototrophic organisms and their indirect impact on the heterotrophic microbial community. As such this can dramatically affect the outcome of the study and the persistence assessment for some compounds. As described in recent publications, there is a simple opportunity to modify the study design by incubation under a light-dark cycle using the same light source as used for algal ecotoxicology studies (which would prevent photodegradation). Inclusion of such a study design as an optional test alongside the irradiated water-sediment study would allow a more complete understanding of the aquatic fate and persistence of many chemicals
- Agree this would be a useful addition.

Disagreement justifications

- Neither agree nor disagree: In our opinion, this would be a completely new test.

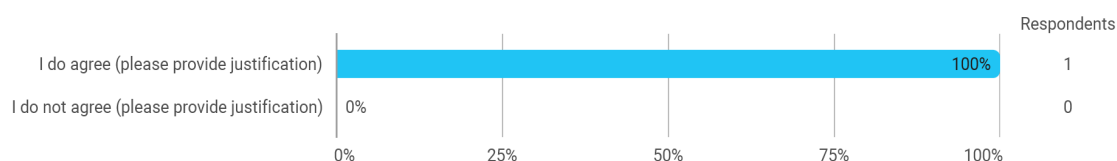
16. para 34 - Include in this section recommendations for substances with particular physicochemical properties, similar to OECD 309 (see paragraphs 22 and 23 of OECD 309).



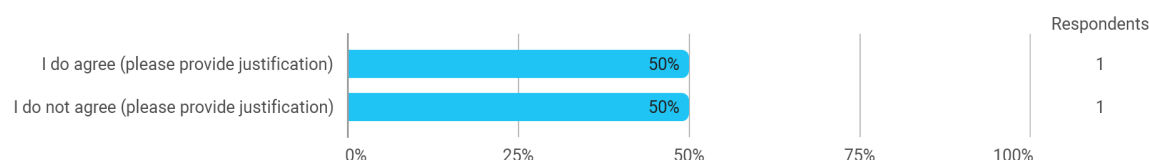
Agreement justifications

- We agree, but unsure whether the relevant science-based advice sufficiently trustable for inclusion in a TG is available

17. para 34 - Indicate that excessively high test substance concentrations should be avoided. Preferably these should not exceed water solubility if possible.



18. para 35 - For substances being dosed with organic solvents, include the requirement to strip solvent off to an extremely small quantity so it does not affect DOC of the test water, and keep amount of solvent used to an absolute minimum (similar to OECD 309).



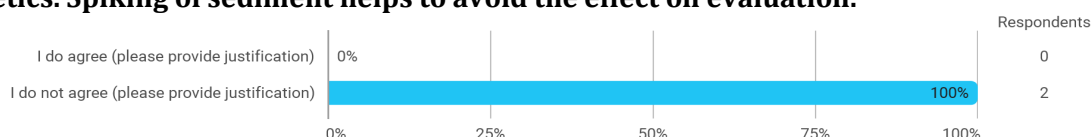
Agreement justifications

- Some solvents are known to deplete quickly the available oxygen, impairing the biodegradation of the test item

Disagreement justifications

- Stripping off solvents without disturbing the sediment or affect the distribution between water and sediment is not possible. Amount of solvent must be minimized

19. para 35 - Add the option to spike the sediment directly with the test substance. Rational: In case a substance is poorly water soluble it tends to adsorb to the glass walls and the sediment thereby complicating kinetic evaluation of the degradation kinetics. Spiking of sediment helps to avoid the effect on evaluation.



Disagreement justifications

- sterile controls should only be considered an optional test as, in most cases, they will not add sufficient value to justify the added complexity and cost of the study
- Further method development and guidance would be required. e.g. it would not be possible to pre-incubate sediment or allow a homogeneous distribution of test substance. Spiking of sediment would affect the distribution of substance in sediment relative to if partitioning from water phase, and affecting the aerobicity of the sediment.

20. para 37 - Request sterile controls

Rational: Sterile controls should be mandatory. They produce important information that often are decisive for interpretation especially if the test substance is difficult to test, e.g. has very low water solubility or is volatile. Clear recommendations how to achieve sterile conditions are lacking and should be added.

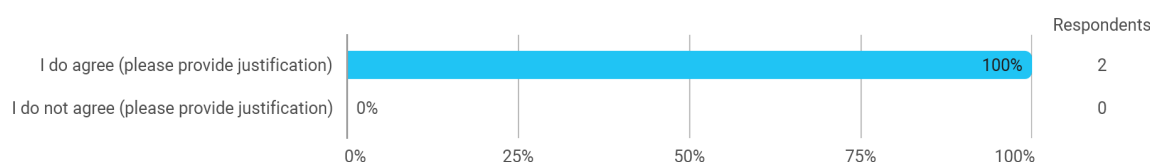


Disagreement justifications

- ▶ Testing in sterile systems should remain optional. The requirement of a sterile system would be a discussion for registration data requirements and not the OECD guideline itself
- ▶ Experience shows that it is challenging to get a true sterile control. We think relevance should be demonstrated with science first, and then included in a regulatory test.
- ▶ Should be included as an optional test only, nevertheless clear advice on sterilisation procedures should be provided, nevertheless.
- ▶ Disagree, sterile controls can be considered optional but should not be a requirement in the OECD guidance (matter for national data requirements)

21. para 38 - Update to reflect newer methods for the determination of the biomass in the control test vessels

Rational: The test guideline is standardized in many ways, but least so with respect to determination of microbial biomass, which is the catalysts of the system. Many of the current methods employed are > 70 years old. There are accurate and precise newer methods routinely used in microbiology, which get closer to measuring the catalytic element of biodegradation in the microbial biomass, enabling standardization/normalization of the determination or providing further information with which to interpret outcomes. For example quantitative real-time polymerase chain reaction (PCR) targeting universal marker genes (Harms et al. 2003), total cell counts using epifluorescence microscopy (EFM) or flow cytometry (FCM). Total cell counts using fluorescent staining of DNA with EFM is traditionally considered the “gold standard” for quantifying bacterial cells in the aqueous environment (Brown et al. 2019).

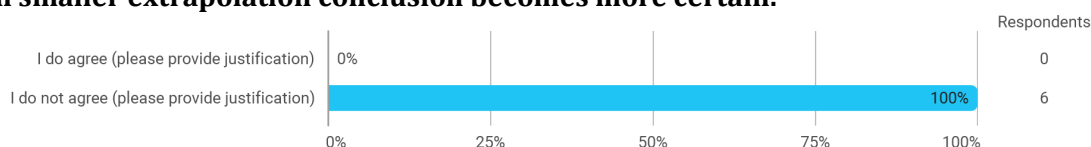


Agreement justifications

- ▶ It would be of value to expand the range of methods beyond SIR or fumigation extraction to encompass newer techniques as options

22. para 38 - Test duration should be set to last at least 120 days. At best 180 days.

Rational: In persistence assessment 120 days is the trigger for persistence and 180 Days the trigger for very persistent substances. Experience shows that the system usually remains stable for 120 days and beyond. The shorter the time frame is to 180 days the less extrapolation is needed or even a direct comparison may be possible. With smaller extrapolation conclusion becomes more certain.

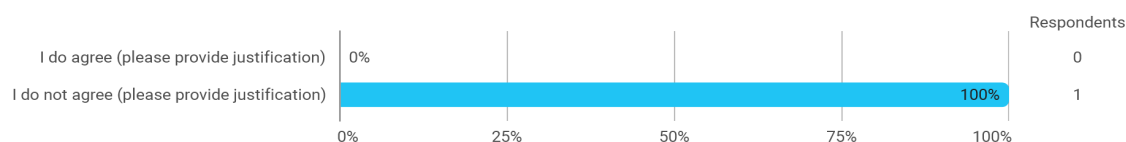


Disagreement justifications

- ▶ Maintaining viability for 100 days can already be challenging (there is often a significant decline in biomass over time). Extending the duration would simply exacerbate this problem
- ▶ There is substantial difficulty in maintaining test system viability for longer incubation periods. Regulatory reviewers frequently reject data collected beyond 120 days due to test system viability concerns
- ▶ very impractical; it is already difficult enough to keep the systems stable and "alive" for 100 days. 120 d would be even more difficult; 180 d is to our experience outside any feasibility.
- ▶ We don't see the need to go for at least 120 days if the relevant information is obtained after 60 days. Extending to at least 120 days could though yield valuable information, e.g. for substances requiring adaptation. No opinion on the possibility to extend to 180 days. It is our understanding that the viability of the inoculum is not guaranteed. This could be temperature dependent.
- ▶ Due to test system viability issues this proposal should not be considered.
- ▶ Extended incubation compromises viability of the test system and can result in erroneous conclusions of biphasic behaviour.

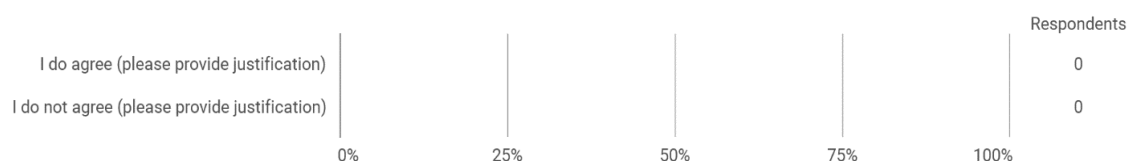
23. para 38 - Update to reflect newer methods for the determination of the biomass in the control test vessels

Rational: The test guideline is standardized in many ways, but least so with respect to determination of microbial biomass, which is the catalysts of the system. Many of the current methods employed are > 70 years old. There are accurate and precise newer methods routinely used in microbiology, which get closer to measuring the catalytic element of biodegradation in the microbial biomass, enabling standardization/normalization of the determination or providing further information with which to interpret outcomes. For example quantitative real-time polymerase chain reaction (PCR) targeting universal marker genes (Harms et al. 2003), total cell counts using epifluorescence microscopy (EFM) or flow cytometry (FCM). Total cell counts using fluorescent staining of DNA with EFM is traditionally considered the "gold standard" for quantifying bacterial cells in the aqueous environment (Brown et al. 2019).

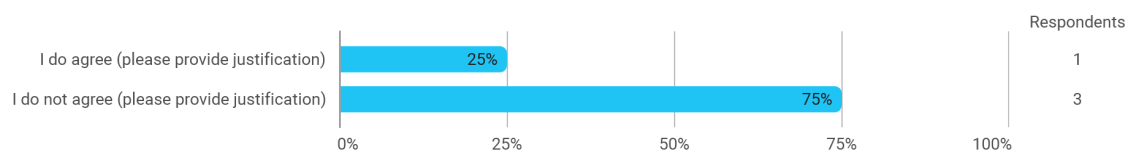


24. para 39 - In addition, the number of sampling points should be enhanced to at least 8 if a substance is prone to quick dissipation because of strong adsorption or volatility.

Rational: Substances that strongly adsorb or are volatile normally dissipate quickly to other compartments. This has consequences for calculation of degradation kinetics and may make it necessary to omit data points. If this is the case and only 6 data points are given it is possible that some data have to be omitted in order to evaluate degradation. Then too few data points remain for calculation of degradation kinetics.



25. para 39 - Extraction of sediment with a proper solvent should be repeated until the extract contains less than 5% of the first extract.



Agreement justifications

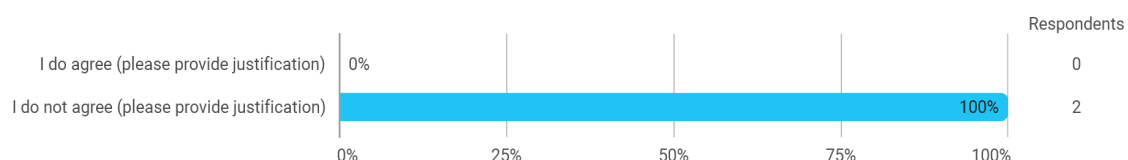
- Extraction efficiency needs to be explained in more detail

Disagreement justifications

- The choice of appropriate solvent will be compound dependent and it will not always be possible to extract >90% in the single first step. The key objective should be to extract > 90% in the 0 day samples with the entire extraction regime, which then establishes this has an efficient method for the remainder of the sampling points
- Reasoning in percentage can be tricky. If the first extraction yields 1%, it is questionable whether a second extraction should be performed
- Unsure what is meant by "proper solvent". Generally extraction is continued until the extract contains less than 5% applied not 5% of the first extract.

26. para 39 and 42 - Detection of volatile substances (parent and metabolites) should be a mandatory request and its relevance emphasised.

Rational: Overall recovery is heavily influenced by volatility of parent as well as transformation products. It impacts study validity and thus efforts to detect volatiles should be mandatory and more emphasised

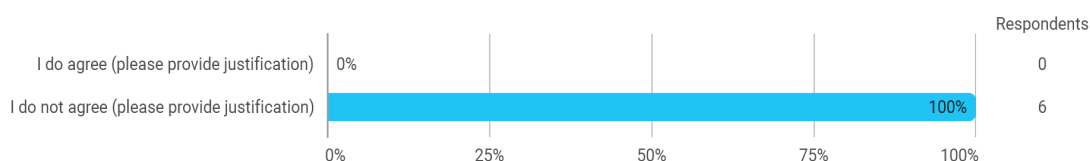


Disagreement justifications

- Volatile parents should not be tested. In addition, identification of degradation products should be optional, depending on the objectives of the study.
- Detection of volatile substances necessary only in certain circumstances (i.e. route of degradation studies)

27. para 42 - Besides quantification NER should be characterized. See Schäffer et al. (2018), ESEU 30:51, to differentiate the three NER types

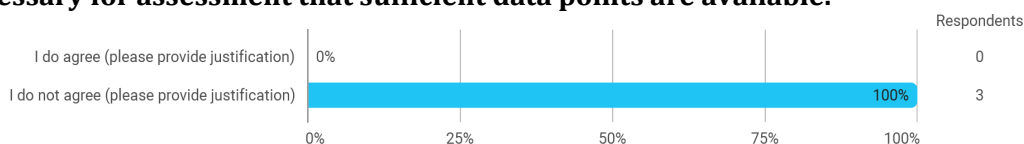
Rational: Type 1 NER is of environmental relevance because it is slowly released from the sediment.



Disagreement justifications

- ▶ Interpretation of NER and its significance varies from region to region. The methods proposed by Schaffer address specific EU interpretation. It is therefore not appropriate for this to be included in a globally harmonized guideline. This can be addressed by separate regional guidance (as the US EPA has done with their 2014 extraction efficiency guidance). Furthermore the proposed methods are not yet fully validated or suitable for many laboratories.
- ▶ this would over complicate the conduct of the study for many regions outside of EU that rely on OECD guidelines. Also, the technical feasibility of the Schaffer et al. proposal has not been fully realized yet. This proposal is still under stakeholder review
- ▶ this concerns the analytical method and has nothing to do with the guideline dealing with the test setup; the analytical method depends on the data requirements of the legislation under which the test item is investigated
- ▶ Strongly disagree. The science is not ready for this. A ring-test is needed before inclusion of such a heavy protocol in a TG. In addition, we are not aware of any published study on NERs in an OECD 308 test to date. We don't think we should transpose OECD 307 to OECD 308 without caution.
- ▶ Although scientifically reasonable, the need for NER characterization should be optional and left to the data requirements of the legislation under which the study has to be performed. The current approach of NER characterization is rather EU centered.
- ▶ Disagree, characterisation of NER is a matter for national data requirements (for example the US EPA guidance on the subject)

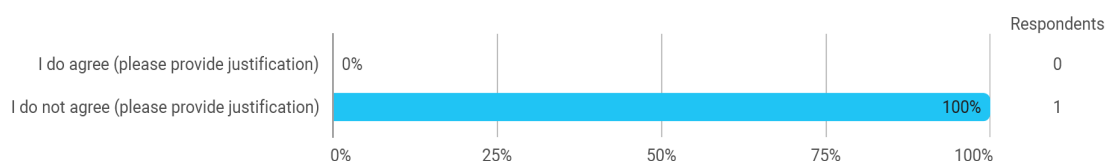
28. para 45 - Sampling Points are too few in case of difficult to test substances
Rational: In case of quickly dissipating substances with slow degradation it is necessary for assessment that sufficient data points are available.



Disagreement justifications

- ▶ The number of data points is not restricted by the guideline
- ▶ Neither agree nor disagree: §45 does not deal with the number of sampling points
- ▶ May be true, but additional sampling points can always be added where required.

29. Annex 1: Annex 1 suggests the levels of oxygen concentration in water for aerobic conditions range from 7 to 10 mg/L. The range should be modified to 6 - 10 mg/L. In situ oxygen measurement of pond water at the field site is usually at about 6 mg/L.

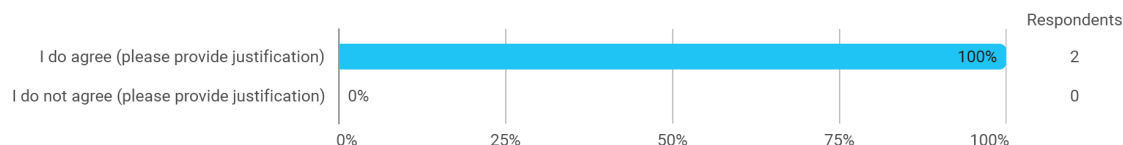


Disagreement justifications

- Oxygen concentration depends on various parameters. If we take temperature as the only influencing parameter, a range of 7-10 mg/L is appropriate.

30. Annex 1. The requirement for the anaerobic test system to be < - 100 mV (Eh) is too restrictive and difficult to achieve. There is literature showing that anaerobicity is induced at higher Eh values than this. Furthermore, the US EPA has issued more nuanced guidance that they will accept studies where Eh7 (corrected to pH7) is < 296 mV and pE (Eh/59.2) + pH is < 12. These alternative criteria for anaerobicity should be added to the guideline

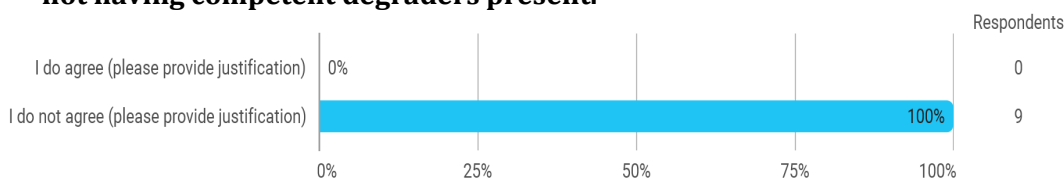
Rational: The requirement for such extreme anaerobic conditions is not justifiable and can lead to rejection of well conducted studies simply on the basis of an arbitrary trigger for a parameter which is extremely difficult to measure accurately.



Agreement justifications

- There are many ways to define anaerobic conditions and an Eh of -100 mV is the most extreme (and very difficult to achieve). A broader definition of anaerobicity would enhance compliance without compromising environmental realism
- Agree, it would be useful to look again at what classes as an anaerobic test system.

31. Increase the biomass to increase the likelihood of degraders being present. Consider potential for inclusion of a series of vessels in study design with increased biomass (alongside standard biomass) to reduce the biological lottery of not having competent degraders present.



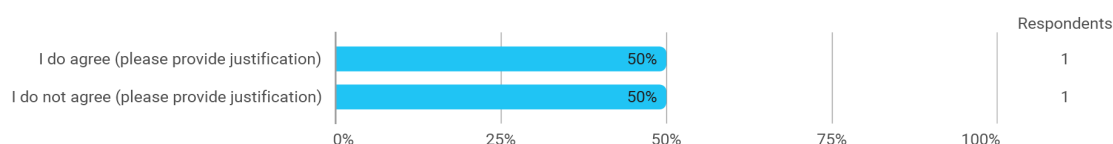
Disagreement justifications

- OECD 308 does not contain too little biomass; this may be true for other types of studies, however not for OECD 308, where the sediment provides for very high numbers of microorganisms. OECD 308 has been shown to be robust test system (Seller et al., 2020, Seller et al., 2021); low inter-replicate variability underlines that.
- Seller et al. (2020 (<https://pubs.acs.org/doi/10.1021/acs.estlett.0c00725>)) state that OECD 308 is a robust test. For OECD 309 high fluctuations and increased variability in between replicates was observed. This is not true for OECD 308, which contains orders of magnitude more microorganisms than OECD 309.

- ▶ The studies are intended to provide information on degradation in realistic test systems, therefore manipulation of the test systems by increasing degrader numbers deviates from this concept, In cases where the behaviour is changed in the manipulated systems it is then impossible to relate this back to the real situation, therefore this would simply add confusion to the interpretation
- ▶ Suggestion would unnecessarily complicate the study design towards a more academic exercise. Augmenting soils with additional biomass could cause the study soils to become no longer representative of the natural soils encountered by crop protection chemicals. This would reduce the usefulness of this study in environmental risk assessments.
- ▶ should be optional; for simulation tests performed for environmental risk assessment, alterations concerning natural biomass will produce artificial results
- ▶ Neither agree nor disagree: This might be difficult to achieve, and the test would no longer be a simulation test.
- ▶ Don't see how this can be achieved in water sediment systems. These systems are normally quite biologically active.
- ▶ On the one hand, there is unclarity regarding the implementation. A precise specification on the implementation of systems with increased biomass should be given. On the other hand, test systems should be environmentally relevant (simulation test) and any manipulation of the test system to increase biomass moves away from this concept and reduces realism of the study.
- ▶ Disagree, the study is designed to simulate a natural water body so this proposal is inappropriate. "Biological lottery" aspect is addressed by testing multiple natural systems (the number and type of which being subject to national data requirements).

32. Additional reference compounds in both screening (301 series, 302) and simulations (307,308,309) tests should be included.

These additional reference compounds should allow to reflect degradability of more complex compounds ((bio)polymers).



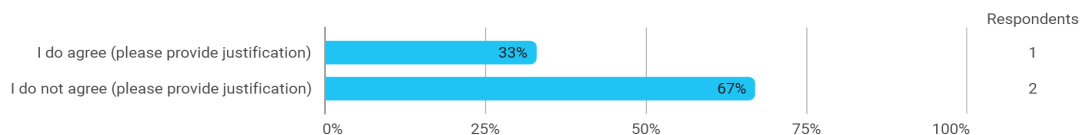
Agreement justifications

- ▶ We agree. The TG lacks relevant validity criteria.

Disagreement justifications

- ▶ Reference substances (in this context) not required for the 308.

33. Quality Criteria - Suggest adding a section on oxygen level, with a phrase like "In case the oxygen concentration drops below 7 mg/L in the water phase, the results should be examined carefully to ensure lack of oxygen is not inhibiting biodegradation"



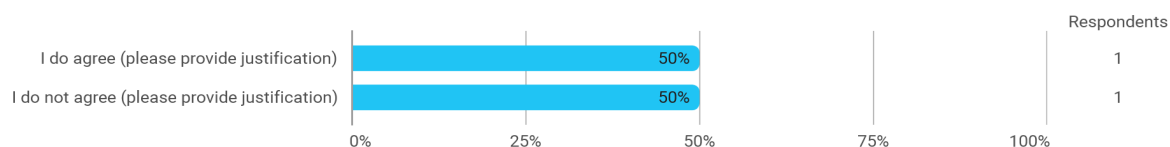
Agreement justifications

- Fully agree. It should be a validity criterion.

Disagreement justifications

- This cannot be included as a general requirement, as OECD 308 has to cover natural variability in a variety of factors influencing biotransformation. One important factor is the redox conditions. In natural environments anaerobic redox conditions are not uncommon. Furthermore, OECD 308 names a combination of conditions (aerobic/anaerobic, high/low OC) to cover different environmentally relevant conditions, to not only measure under best case biotransformation conditions but to get an idea of the range of biotransformation found in the environment. The aquatic environment is not a standardized thing and natural variability has to be covered.
- Establish how low oxygen levels typically decrease to, and at what point degradation is expected to be impacted.

34. Validity criteria - The test should include validity criteria. Quality criteria are not enough because they are used to assess the validity of a study. One criterion could be that the test substance has stayed in the sediment/water system. If not, the test is not valid. Also, the test should also give some range of acceptable degradation kinetics of the reference in function of the test temperature. Oxygen in water should not fall below 7 mg/L.



Agreement justifications

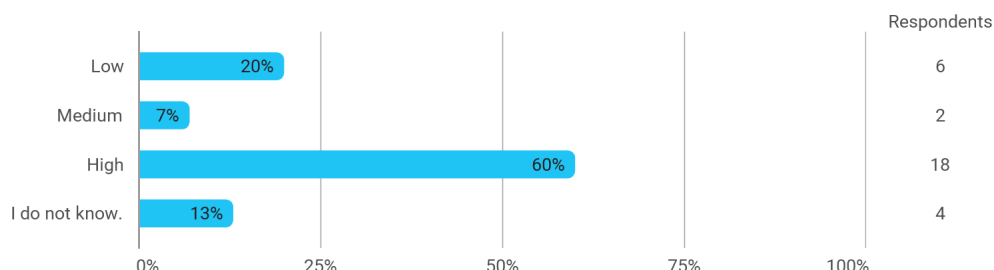
- Fully agree. It should be a validity criterion.

Disagreement justifications

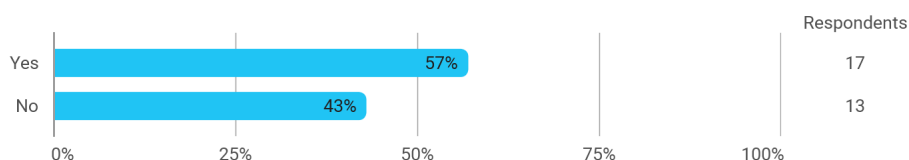
- Repetition of "Suggestion 1" - see above.

A.21 Test No. 309: Aerobic Mineralisation in Surface Water – Simulation Biodegradation Test

Relevance of revision



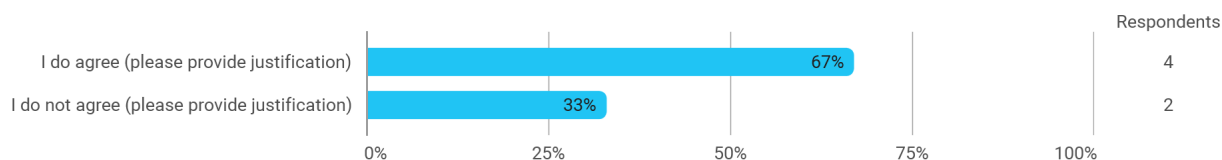
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

- The test should include more validity criteria. The test substance, if not degraded, should be proven to have stayed in the water phase throughout the test (10% loss at the end of the test allowed max). If not, the test should be considered invalid. Also, the test should also give some range of acceptable degradation kinetics of the reference item in function of the test temperature. Oxygen concentration in water should not fall below 7 mg/L.**

Rational: It is impossible to investigate degradability when the substance leaves the water phase by sorption or evaporation. In that case, the study should not be considered valid because it is not fit for purpose. A more reliable and clear validity criterion is required for the reference substance. The oxygen validity criterion is of utmost importance to ensure that the test accurately reflects an aerobic DT50, which is the purpose of the test



Agreement justifications

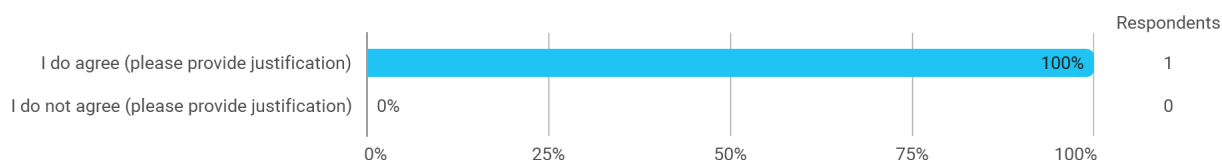
- ▶ Validity criteria are needed
- ▶ There are many validity criteria in the guideline (e.g. mass balance criteria for 14C studies). However, it would be helpful to have more definition of the acceptable degradation of the reference control
- ▶ this test is not suitable for lipophilic substances, because compound will adsorb to the test vessel wall
- ▶ Fully agree. Furthermore, O2 concentration should be a validity criterion.

Disagreement justifications

- Agreed regarding the range of acceptable degradation rate of the reference substance. According to current OECD TG 309 an additional validity criterion is already given, being an acceptable recovery range of 90 to 110% (radiolabelled test substances) and 70 to 110% (unlabelled test substances) of the applied amount.
- Validity criteria already in place (90 to 110% for radiolabelled studies; 70 to 110% for non-radiolabelled studies)

2. Introduce guidance on the best technique(s) to prepare the sterile control (309, 307 and if introduced for 308).

Rational: It is proposed to review if the list of techniques for preparation of sterile controls provided in the simulation TGs need to be updated.

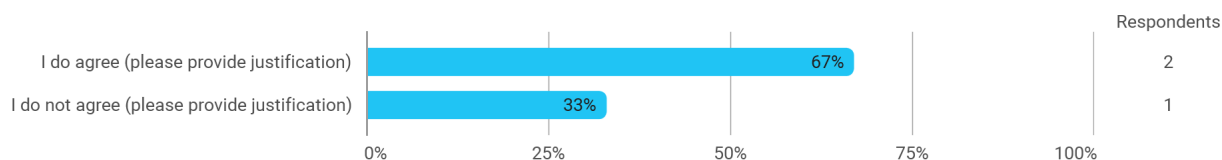


Agreement justifications

- This should be introduced in the TG. It is a bit excessive that some authorities are asking for gamma-ray sterilization while this has never been fully tested nor developed as SOPs in CROs.

3. Clarify if and how simulation and bioaccumulation (this was missed from the survey form for bioaccumulation [OECD TG 305] test) tests for substances containing more than one constituent (multi-constituents, UVCBs etc.) could be feasible/conducted. Reflect current analytical capabilities.

Rational: Simulation tests are not clearly limited to the 'pure' substances (OECD TG 305 even notes that test might be applicable for multi-constituent substances), but testing of complex chemicals is limited by the objective of the test methods (i.e. endpoints investigated) and capability of analytical methods. It is proposed to further clarify if and how complex chemicals could be tested in simulation and bioaccumulation studies to obtain meaningful results. Some review of analytical methods and experiences available is provided in the published report https://echa.europa.eu/documents/10162/13630/pfab_750_06_wp4_echa_final_report_en.pdf/b3a7e562-bf9c-ef02-948f-eaf1b8f89e3f.



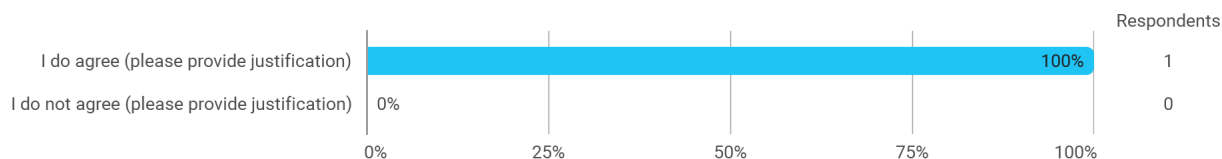
Agreement justifications

- We agree, but this may need further research.
- Important aspect as OECD 309 has been designed for monoconstituent test substances, not complex ones like MCSs and UVCBs

Disagreement justifications

- No strict explanation possible although I would prefer to have such one.

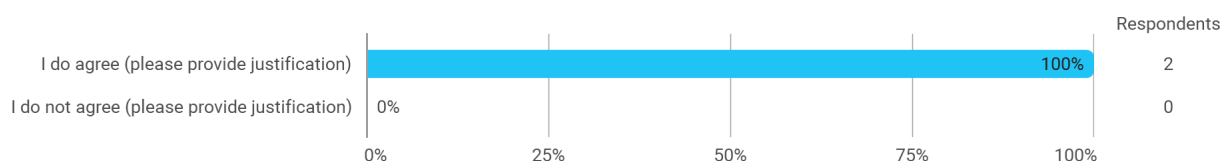
4. para 1, last sentence: This is not necessarily necessary with today's equipment. Might delete this sentence.



Agreement justifications

- This is sometimes necessary even with today's equipment

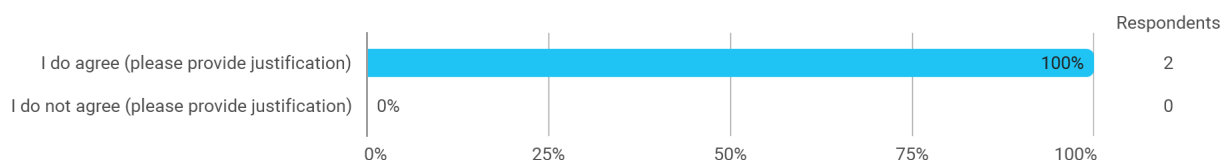
5. para 1 - Suggest to add in the introduction a paragraph explaining what the domain of applicability for this test guideline is (which types of substances it works best for)



Agreement justifications

- We suggest to also clearly specify what kind of substances it does not work for
- Important aspect as OECD 309 has been designed for monoconstituent test substances, not complex ones like MCSs and UVCBs

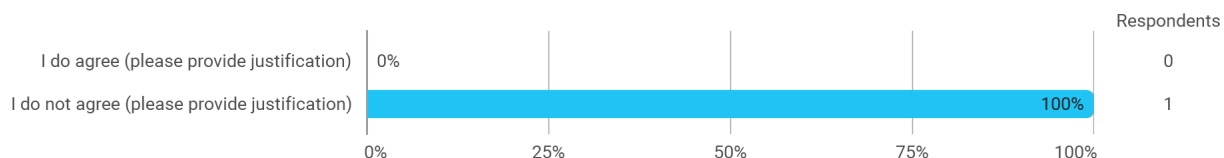
6. para 1 - Suggest adding clear cut-off for Henry's constant, Koc, water solubility, etc, for which this test guideline does not apply. This could build on the properties of the substances on which this guideline was successfully vs unsuccessfully applied until now.



Agreement justifications

- Too many unfeasible and irrelevant tests are requested. There is no point trying to force a substance to stay in water (or sediment) while it clearly won't. It is of paramount importance to give clear phys-chem boundaries.
 - Agree as registrants often experience challenges with decisions from authorities which might be in conflict with the "real life" feasibility of the current guideline
- ### 7. para 2, lines 3-11: Based on results of Seller et al., 2020, we now know that observation of first-order kinetics in OECD 309 studies is very unlikely. We therefore suggest getting rid of all these speculations on lines 3-11 of this paragraph. This is even more so since later a two-phase model that includes a lag phase is recommended

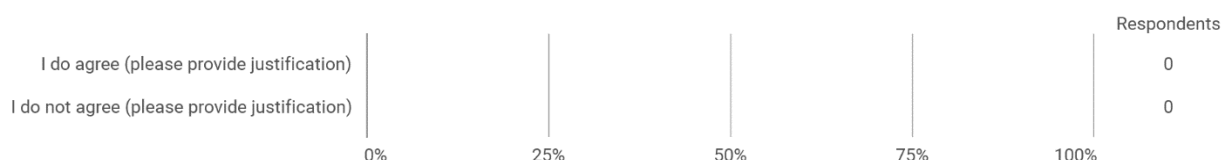
for evaluation of the kinetic data. This would be needed if actual first-order behavior was observed.



Agreement justifications

- Reading Seller et al. (2020), it seems no true OECD 309 was conducted (suspended solids between 1 and 10 g/L). In the EU, suspended solids is typically between 10 and 20 mg/L. This may affect the kind of kinetics observed. Suggestion n°7 could perhaps be refined.

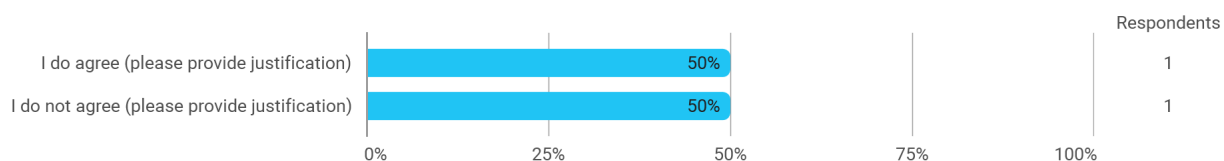
8. para 2, lines 12-14: But actually, mostly not observed.



9. para 2, lines 13-15: and if the degrader biomass changes significantly during the experiment (which frequently happens)



10. para 1 – The guideline contains a sentence "However, an optional secondary objective of the test is to obtain information on the primary degradation and the formation of major transformation products". It is suggested to consider making it mandatory also in OECD TG 309 to measure primary degradation and transformation products, unless it can be justified that only the measurement of mineralisation is sufficient for a particular purpose.



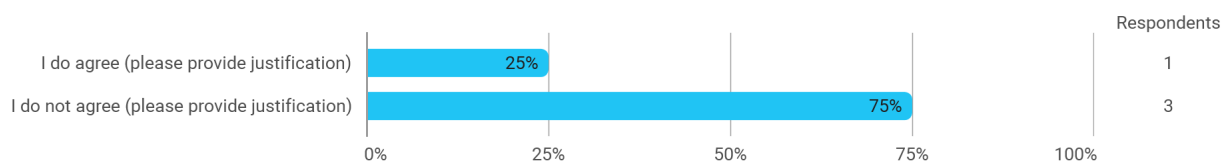
Agreement justifications

- justified that only the measurement of mineralisation is sufficient

Disagreement justifications

- We disagree. These are regulatory considerations, which should be treated at the regulation level, while OECD TG should stay at the recommendation level.

11. para 5, first sentence: Suggest changing the range to 0.1 to 10 g/L (based on Seller et al., 2019) and give suggestions for what purpose which sediment concentrations are most suitable.



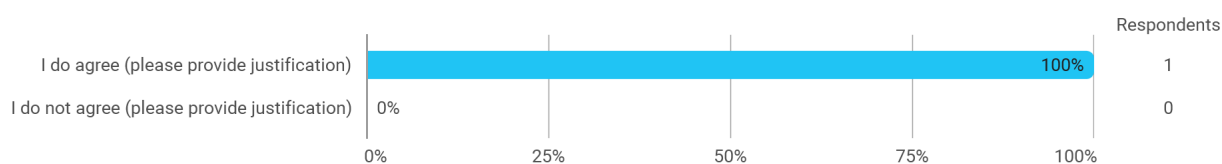
Agreement justifications

- if possible and meaningful

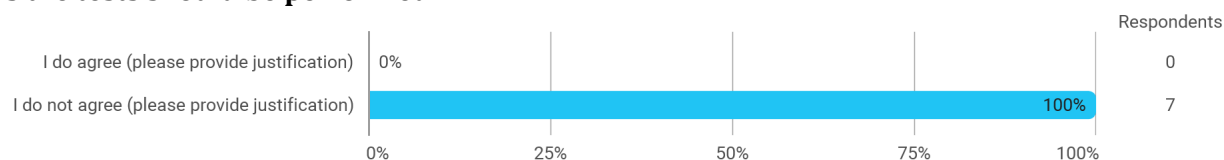
Disagreement justifications

- The current maximum suspended sediment concentration (1 g/L) is already very high with respect to large, open water bodies (which is the stated protection goal for this study in the EU pesticide data requirements) and, as such, applicants for such registrations will typically use a pelagic test or use the lower sediment concentrations. Increase as proposed reduces the discrimination between the OECD309 and 308 test designs
- Neither agree nor disagree: This comment deserves more explanations. It's important to remember that some authorities request SS between 0.01 and 0.02 g/L
- A concentration range ranging up to 10 g/L moves away from the intention to represent open water bodies. At a concentration of 10 g/L the sediment is hardly to be kept in suspension.

12. para 5, line 7: Add a sentence here which says that shaking rather than stirring with a magnetic stirrer should be used (based on Seller et al., 2021, and Shrestha et al., 2016).



13. para 11: At least PEC should always be estimated to decide at which concentration levels the tests should be performed.



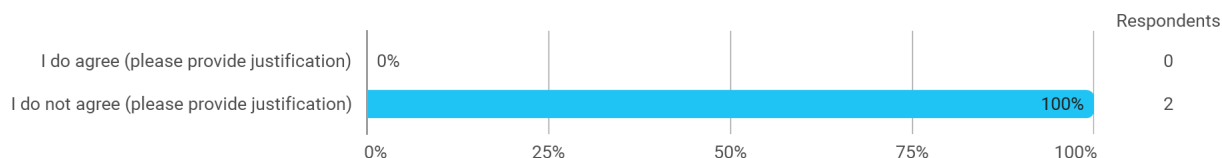
Disagreement justifications

- PECs (if available at the time of study conduct) will often be too low for analytical sensitivity, especially for pesticides with low application rates and the requirement to identify metabolites at > 5% applied radioactivity. It will likely be impossible to conduct the test and an appropriate PEC therefore the current stated test concentrations are acceptable
- PECsw can vary substantially depending on intended use patterns and environmental scenarios. It can happen that at the time of study initiation, not all necessary information for

a PEC estimation is yet available. Given how contentious PEC calculation can be in regulatory submissions and how frequently country specific modelling procedures are revised, the PECsw is not a reliable method for establishing a dose concentration.

- ▶ That should remain optional. It would not be convenient for all kinds of purpose. It should stay up to the regulation to prescribe the conditions of the test
- ▶ Since PECsw can vary depending on environmental and use scenarios. At study start, in many cases not all the information is available that is needed to calculate a reliable PECsw. In addition, calculated low PECsw values used as test concentrations could result in analytical challenges.
- ▶ It is rather “PECs” in case of different identified uses (e.g EU REACH). Such exposure information should not be mandatory to perform a study, which is part of hazard assessment

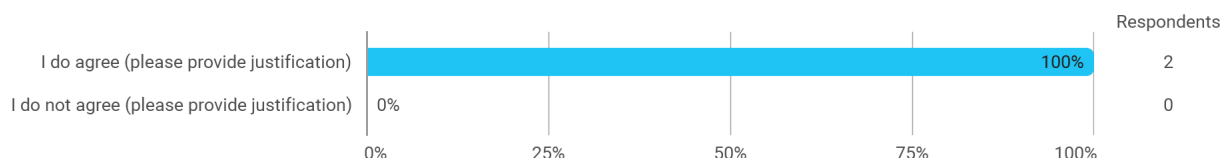
14. para 12: Remove aniline because it readily forms NER with many compounds. Add some other options for activity control, e.g., atenolol for primary biotransformation. Further, today there are many methods available to test for microbial activity which would not require to use reference substances and to perform additional test with those prior to the actual biotransformation test. If reference substances are used, testing the community’s ability for biodegrading chemicals should be assessed in parallel and not prior the actual transformation test. Otherwise, the longer storing times for the inoculum used for the actual test may lead to changes in the inoculum relative to the inoculum run with the reference chemicals.



Disagreement justifications

- ▶ bioNER? Never observed NER! Reference substance is always tested in parallel, not prior to the test!
- ▶ Neither agree nor disagree: Unclear whether sufficient data is available to include this proposal in standardized regulatory testing

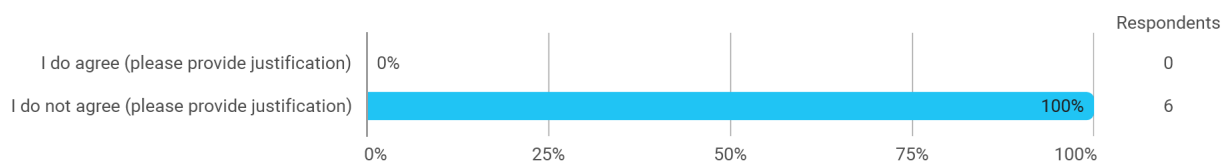
15. para 12 - Reiterate, as was done in OECD 301 guideline, that some reference substances may biodegrade without an inoculum.



Agreement justifications

- ▶ Instead, propose less easily biodegradable substances as reference
- ▶ Very relevant aspect: such reference substances may not be a realistic representation of the microbial viability of a test system.

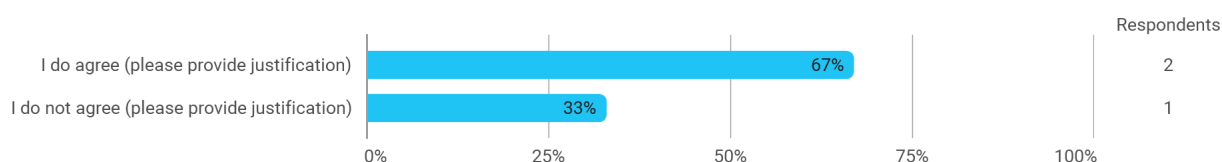
16. para 13, first sentence: Would recommend using a larger number of replicates as Seller et al. (2020) reported large inter-replicate variabilities (i.e., intrastudy variability= in primary degradation of pharmaceuticals and pesticides).



Disagreement justifications

- For radiolabelled substances duplicate samples are sufficient and over the time-course of a study with up to 8 sampling points, outliers can be easily identified. More replicates may be helpful for non-radiolabelled compounds, but this should be expressed as a suggestion not a requirement.
- The number of replicates are sufficient when radiolabelled test substance is used. Additionally, since a time series of data points is analyzed, outliers are easily identified and can be addressed.
- more replicates should be optional; for ¹⁴C-labelled test items, 2 replicates are sufficient for detecting outliers
- Neither agree nor disagree: Seller et al. (2020) did not investigate the true pelagic OECD 309. There is a risk of generalizing too much their findings.
- A larger number of replicates may be useful when applying a non-labeled test item. In case of a radio-labeled test item, the analytical variability is lower and 2 replicates should be sufficient. Also, degradation is measured at several sampling intervals and the therefore a chance to identify outliers is given.
- Increased number of replicates may be useful in certain circumstances, but should not be part of the guidance.

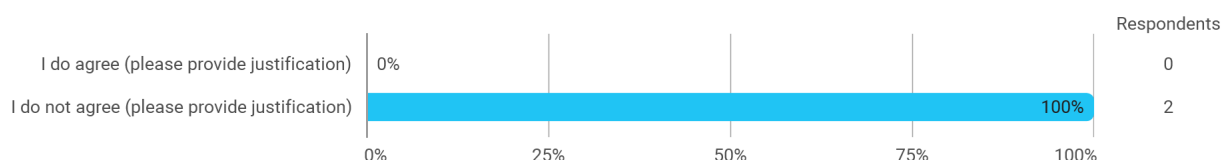
17. para 13 - Add a request to measure and report extraction efficiency and give a Definition.



Agreement justifications

- definition needed
- This would be welcome, indeed.

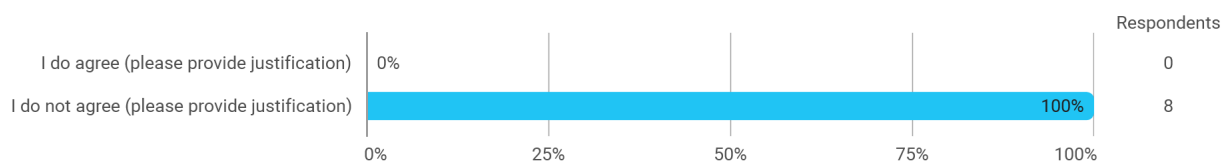
18. para 15, last sentence: Not with today's instruments!



Disagreement justifications

- We don't find this suggestion specific enough.

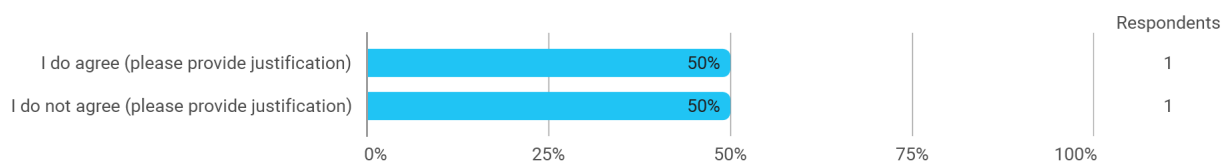
19. para 16, lines 3-4: Tests should be run at least in triplicate because intrastudy variability has been observed to be high, particularly for low or no sediment (Seller et al., 2020)



Disagreement justifications

- Even with 3 variability could occur - should than be done with each degradation simulation OECD test
- See previous. Duplicate is sufficient for 14C studies
- The number of replicates are sufficient when radiolabelled test substance is used. Additionally, since a time series of data points is analysed, outliers are easily identified and can be addressed
- more replicates should be optional; for 14C-labelled test items, 2 replicates are sufficient for detecting outliers
- Depends on the substance, for most substances we observed low variability between replicates. To improve kinetic calculations, It is advisable to increase the number of samplings instead of number of replicates
- Seller et al. (2020) investigated water-sediment systems only.
- A larger number of replicates may be useful when applying a non-labeled test item. In case of a radio-labeled test item, the analytical variability is lower and 2 replicates should be sufficient. Also, degradation is measured at several sampling intervals and the therefore a chance to identify outliers is given.

20. para 16, first bullet point: Delete "or magnetic stirrer"



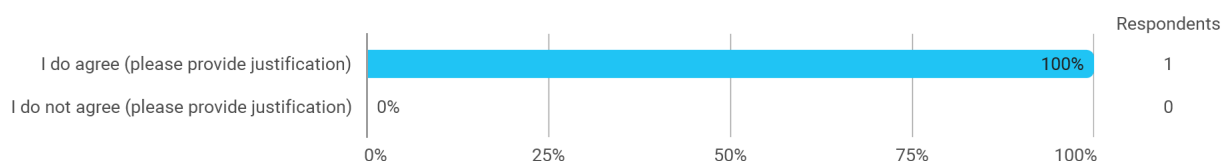
Agreement justifications

- Agree for typo

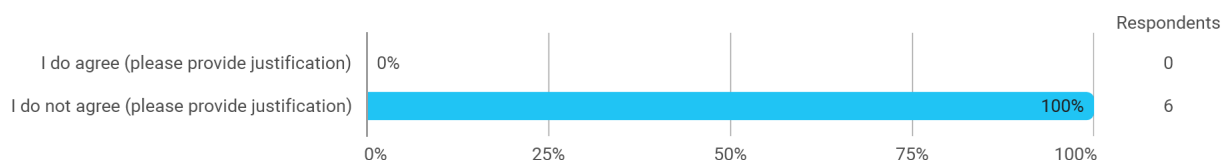
Disagreement justifications

- Unclear why this should be removed.

21. para 16, ninth bullet point: Add text to clarify that sediment sampling and analysis should be a must for non-pelagic tests.



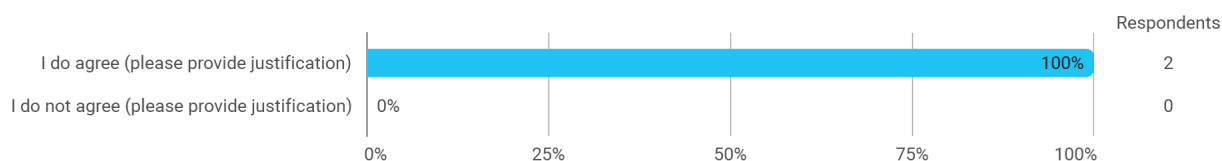
22. para 16, last bullet point: Flow cytometer, if possible, for determination and characterisation of biomass.



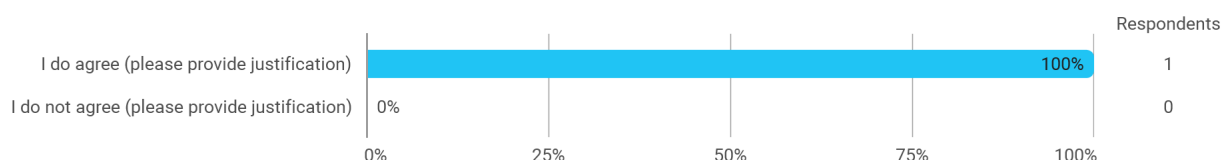
Disagreement justifications

- ▶ Such methodologies are not readily available in most laboratories conducting these studies
- ▶ The use of flow cytometry should be optional, it is not standard equipment at the CRO's performing this test. Adequate methods for determining microbial biomass are already used.
- ▶ flow cytometer should be optional; it is a good as any other "biomass determination"
- ▶ Overdone!
- ▶ Should only be optional, not a requirement.
- ▶ Flow cytometer should remain optional as this will not be available in many labs.

23. para 17 - Merge paragraph 17 with paragraph 22, rather than keeping them separate.



24. para 18, lines 3-6: Why? Many chemicals will enter "pre-exposed" water bodies when discharged e. g. through WWTPs. Sampling sites should be chosen based on the expected point of entry and distribution into the aquatic environment. Maybe set a minimum distance to WWTP input instead.



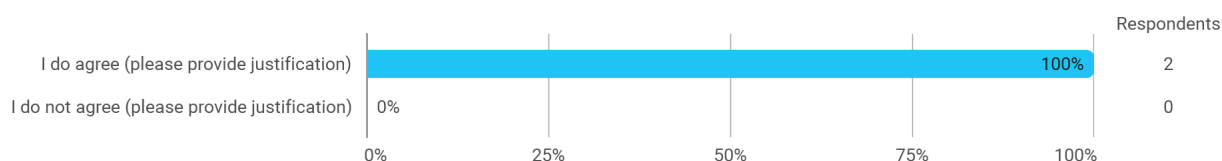
Agreement justifications

- ▶ Relevant comment: could increase real life (environmental) relevance of such tests.

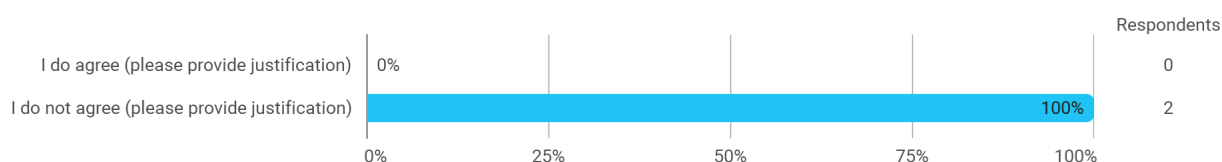
25. para 19 - No pre-equilibration phase is stated. The guideline states that the test should ideally be started within 1 days of sampling, but it is not clear whether starting means

set up of the test vessels or application of the test item. Therefore, some guidance on the gap between set up of test vessels under the test conditions and actual addition of the test item would be beneficial

Rational: Clarification of text to align with other TGs



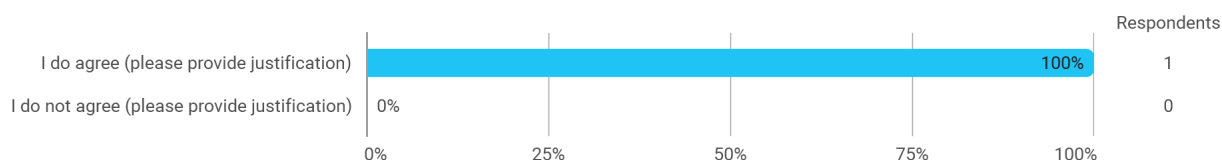
26. para 20, first sentence: Suggest changing to 0.1 to 10 g/L.



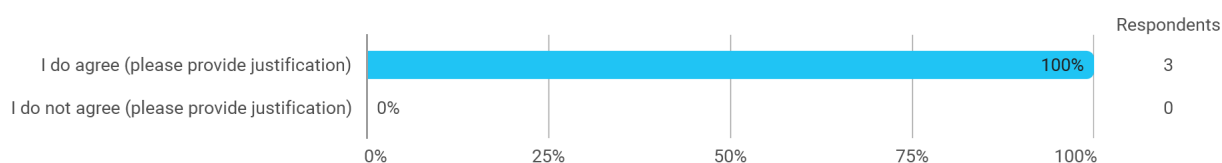
Disagreement justifications

- See previous. Such a high sediment concentration does not represent a large, open water body
- In the EU, suspended solids is typically between 10 and 20 mg/L. It may be an issue to change the TG to the 0.1 ; 10 g/L range

27. para 23 - Add passive dosing as one of the options for dosing of poorly soluble substances.



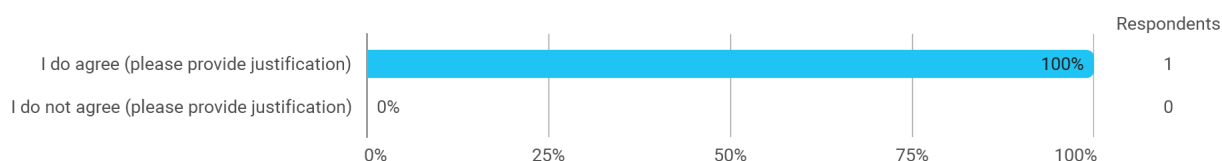
28. para 23 - Add a statement alerting the reader that the use of highly biodegradable solvents can lead to anaerobic conditions in the test system.



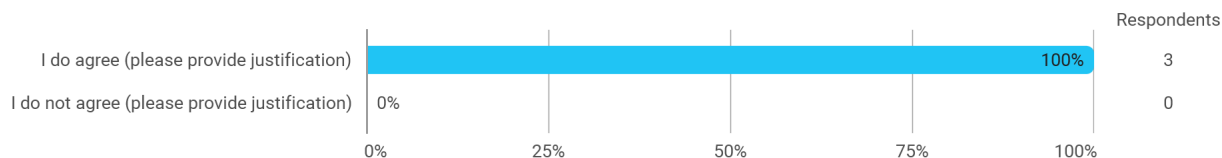
Agreement justifications

- highly biodegradable solvent can impact biodegradation of test substance and should be avoided!
- Fully agree
- agreed

29. para 23 - Add a statement alerting the reader that use of a co-solvent may also lead to formation of a bioavailability-limiting, third phase in the test system.



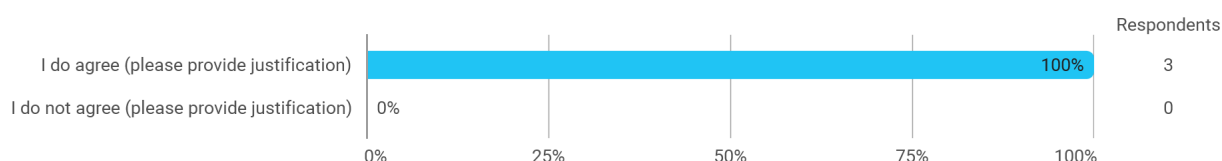
30. para 24 - It should be considered to add advice how the validity criteria should be applied when the incubation temperature differs from that used for deriving the validity criterion standard temperature.



Agreement justifications

- ▶ Yes
- ▶ Fully agree. It is known that temperature (and type of test systems) can affect degradation kinetics of sodium benzoate, sometimes to the point that it almost meets the P criterion of REACH ! True validity criteria are needed for this TG !
- ▶ Fully agree, highly relevant: the differentiation of the validity criteria for e.g. 20° and 12°C is not sufficiently clear.

31. para 24 - More definition of the term “diffuse light” would be beneficial, e.g. specifying that this is non-UV, PAR light (as used for plant growth/algal ecotoxicology studies)
Rational: The use of diffuse light can be a valuable option to encompass the potential for algal metabolism, but more definition is required to ensure consistency of interpretation

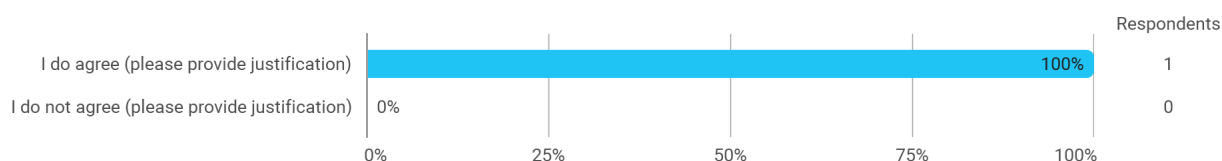


Agreement justifications

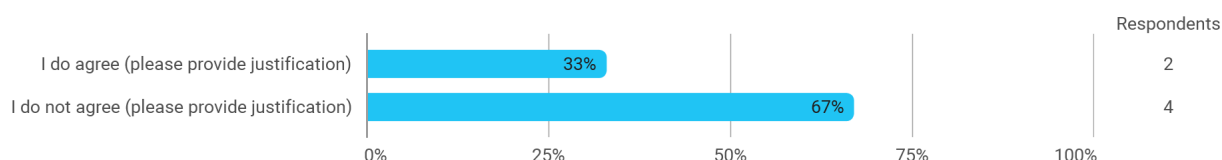
- ▶ Diffuse light should be stated in more detail
- ▶ The use of diffuse light is a valuable addition as it can allow for metabolism of phototrophic organisms (as per much published literature), but it is a vague term and would benefit from a clearer definition
- ▶ Agree, clarification would be useful here.

32. para 25 - Agitation by means of continuous shaking or stirring must be provided to maintain particles and microorganisms in suspension. Agitation also facilitates oxygen transfer from the headspace to the liquid so that aerobic conditions can be adequately maintained. Place the flasks on a shaking table (approx. 100 rpm agitation) or use magnetic stirring. Especially, the use of magnetic stirring is not encouraged while preparing suspended sediment test using sandy sediment. As

any deposition of heavy sand particles results in grinding of sediment and abrasion of magnetic stir bar which can lead to artefacts on the test results. Agitation must be continuous. However, the shaking or stirring should be as gentle as possible, while still maintaining a homogeneous suspension.



33. para 25 - Add in this section "There may be different incubation temperatures required for regulatory application of the test outcome."



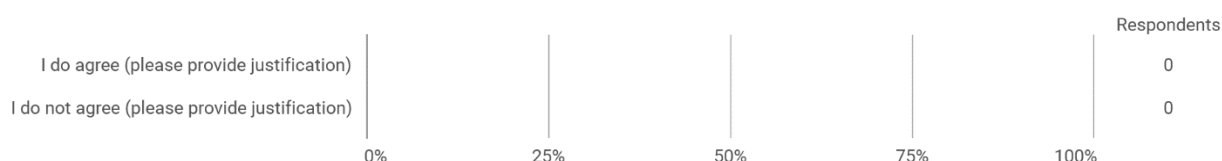
Agreement justifications

- Temperature for e.g. REACH (12°C)
- No specific comment

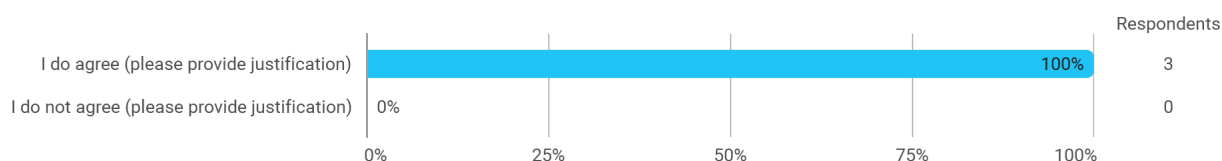
Disagreement justifications

- Differences in degradation related to temperature can be accounted for using accepted normalization methods without the need to perform the study at different temperatures
- Additional testing should not be required to obtain information that can easily be determined via modelling. Default factors and calculation methods for describing temperature response can be used to calculate DT50 / DT90 at different temperatures if required.
- testing at different temp. may become technically quite challenging; however, depending on legislation, default factors and calculation methods for describing temperature response can be used to calculate DT50 / DT90 at different temperatures if required
- Depending on the legislation, default factors and calculation methods can be used to calculate DT50 / DT90 at different temperatures if required.

34. Paragraph 26, last sentence: to ensure the robust estimation



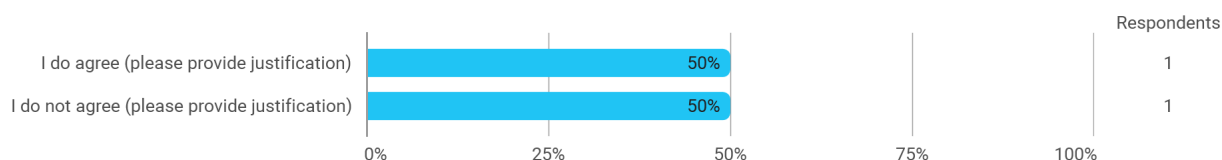
35. Paragraph 28, lines 5-7: It would be nice to add some words what the intention behind using two different test concentrations is.



Agreement justifications

- would be nice!
- Testing of different waters instead of 2 concentrations should be considered
- It would also be good to take into account the limit of detection of the analytical apparatus. For instance, sometimes, the radioactivity of the test item makes it mandatory to start at 50 µg/L in order to be able to follow degradation up to 90%. 50 µg/L may be above the water solubility limit, which is not relevant for such test. Unclear whether the study should be considered not technically feasible in that case. The same would be the case for the concentration to be used to identify the degradation products (e.g., 500 µg/L)

36. Paragraph 30, lines 10-13: Given the previous speculation about SFO kinetics, it could be advisable to distribute the sampling times so that they follow a geometric pattern (e. g. 1,2,4,8, etc. days, adapted to the expected pace of degradation)



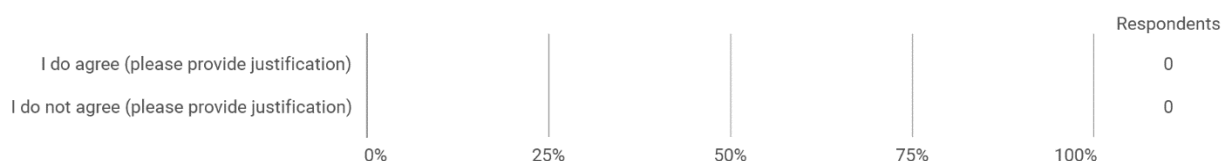
Agreement justifications

- Agree that the sampling times should have a relevant spacing. Providing more advice in that space may be beneficial if possible

Disagreement justifications

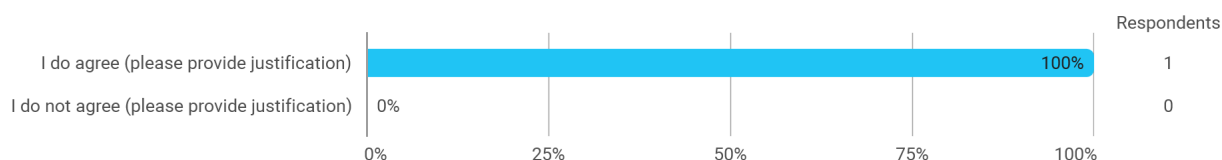
- Optimal distribution of sampling times must be evaluated for each substance in pre-test

37. Paragraph 31, first bullet point: Note that if biomass measurements by flow cytometry are required or planned, flasks need to be planned such that sacrificial sampling is possible for non-pelagic studies, at least at those time points where biomass should be measured (e.g., t0, one intermediate time point, and at the end). Sacrificial sampling of non-pelagic tests is needed for FC analysis because it would be very difficult to draw a representative sub-sample from flasks with suspended sediment.

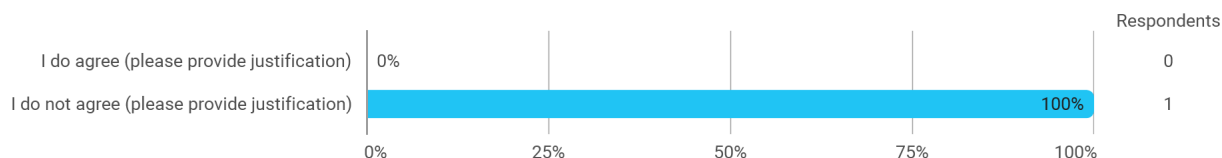


38. Paragraph 31, fourth bullet point: Maybe define explicitly that the test with reference substances has to be run in parallel as we know from practice that usually this is done

before the actual test and therefore not necessarily represents the activity of the inoculum.



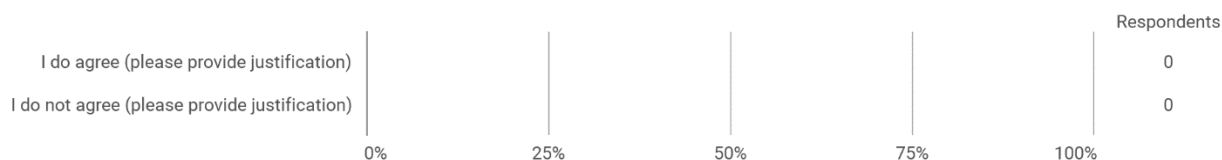
39. Paragraph 44, first sentence: Preparing for a lag phase contradicts the expectation to get SFO kinetics



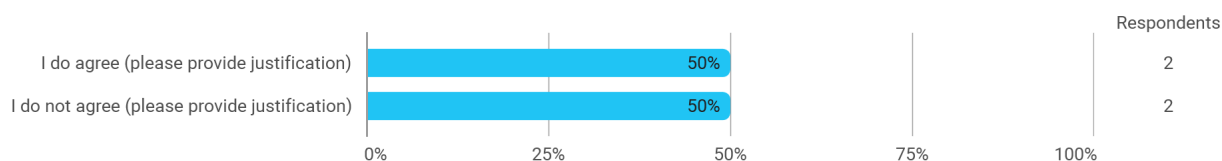
Disagreement justifications

- We disagree, unless the calculated lag phase is removed from the SFO kinetic derivation.

40. Paragraph 44, lines 3-5: This could be done better by fitting the length of the lag phase and k simultaneously



41. Paragraph 44, lines 13-15: Averaging rate constants does not sound to be a good idea. Rather a geometric average, if any.



Agreement justifications

- Although it is not clear what is meant by this averaging (i.e. is it averaging across replicates or averaging of 2 rate constants from biphasic decline), it is the case that the precise method for performing kinetic assessments of all studies is better covered in specific regional guidance designed to reflect specific modelling processes (e.g. FOCUS in the EU)
- although this is not a question of test design, outdated kinetic data evaluations should be discarded

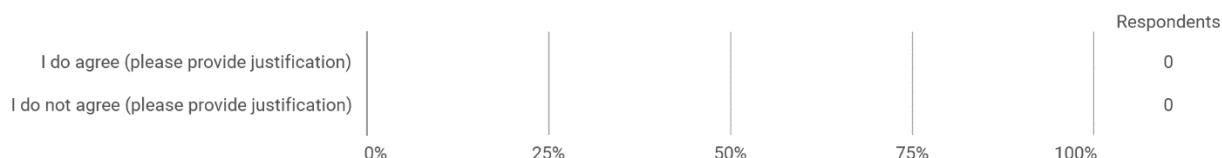
Disagreement justifications

- When following FOCUS kinetics, the fitting is usually done using all replicates of an individual experiment in one fit. This will lead to an “averaged” fitted curve. the fitting statistics provide information on parameter uncertainty (parameter t-test and / or confidence intervals...) If overall mean DT50 / DT90 are required (i.e. averaging across

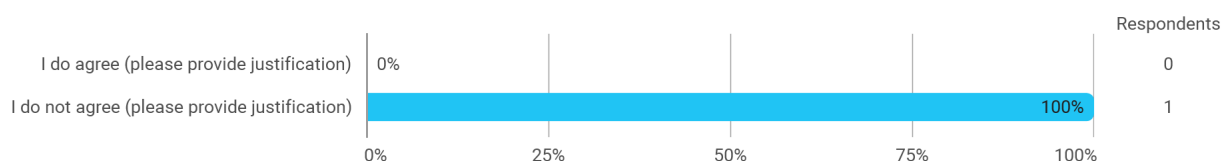
different test systems, test concentrations...), then a geometric mean value should be recommended.

- As data interpretation varies across regions, it should not be part of the OECD test guideline and should be specified by kinetic guidance, applicable for the respective region (e.g., FOCUS in the EU).

42. Paragraph 46: It would be important to emphasise that measuring substance residues in sediment is a must for suspended sediment tests, otherwise it is not possible to distinguish between sorption and degradation.



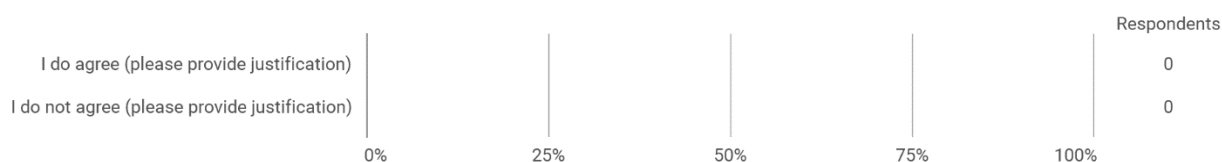
43. Paragraph 47, lines 7-11: This phase may overlap with some NER formation for suspended sediment tests.



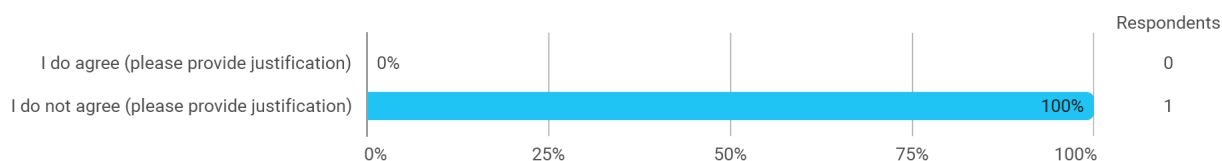
Disagreement justifications

- Neither agree nor disagree: Please note that this TG does not mention NER

44. Paragraph 49, first sentence: This is wrong when we allow for a lag phase for biotransformation. Hydrolysis presumably does not have a lag phase, so the combination of hydrolysis and a lagged biotransformation would result in multi-phase kinetics. (of which the subtraction would only work after the unknown lag time).



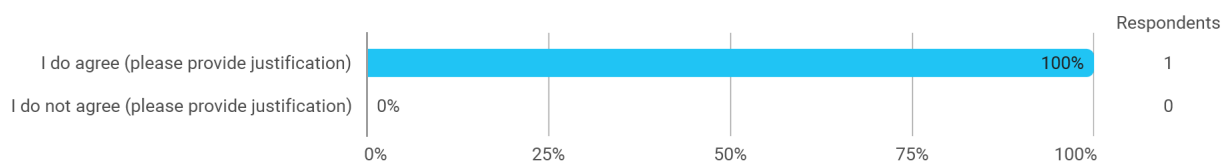
45. Paragraph 50, penultimate sentence: It would be helpful to perform this at many (all) sampling occasions, not only at the end.



Disagreement justifications

- Neither agree nor disagree: Unclear why this would be helpful

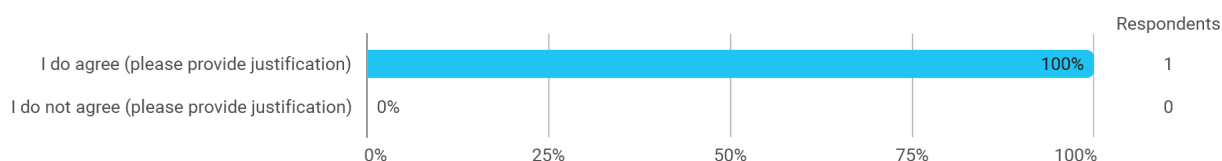
46. Paragraph 51: If the recommendation to remove aniline as reference substance is followed, this whole paragraph should also be deleted. Is there similar data for benzoic acid available that could be cited here instead?



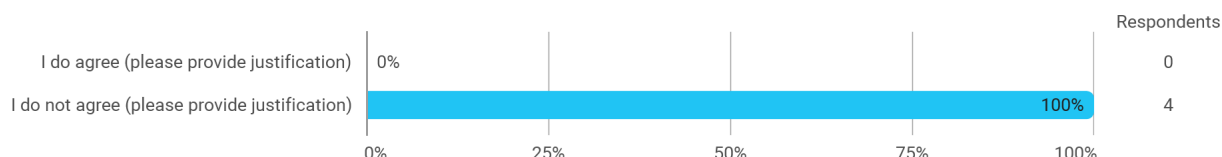
Agreement justifications

- if aniline is removed, I agree

47. Paragraph 51, penultimate sentence: Cell densities in surface water bodies are generally higher (see, e.g., Seller et al., 2020). Reference to the documentation of this ring test should be included.



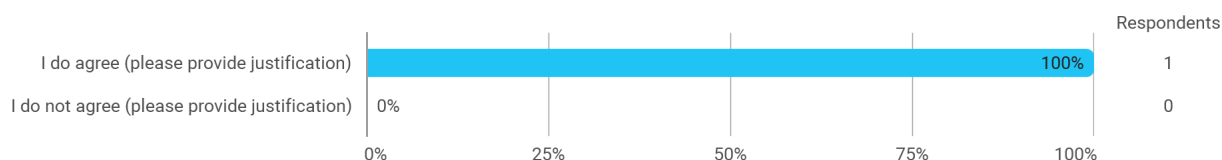
48. Paragraph 53, first bullet point following “and optionally”: This should be made mandatory and a suggestion should be given to use flow cytometry counts for biomass measurements (see Seller et al., 2021, doi: abs/10.1021/acsenvironau.1c00006)



Disagreement justifications

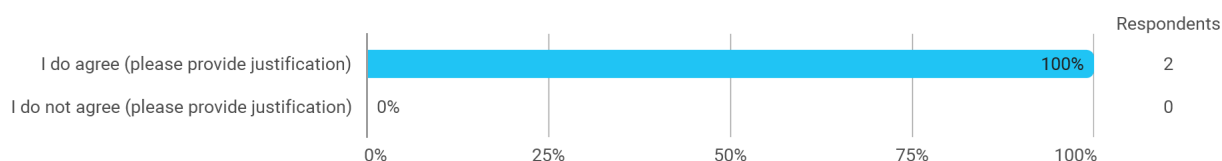
- Flow cytometry is not a common technique in most laboratories conducting these studies, so it should remain optional
- The use of flow cytometry should be optional, it is not standard equipment at the CRO's performing this test. Adequate methods for determining microbial biomass are already used.
- viability of system is ensured by reference substance; flow cytometry is just another "biomass" determination which gives inconclusive numbers
- Neither agree nor disagree: Added value unclear

49. Paragraph 53, third bullet point following “Results”: including sediment-bound aR% for suspended sediment tests



50. para 51 - The biodegradation of sodium benzoate and aniline may not be a realistic representation of the microbial viability of a test system. The extent (complete? Pass level of 60%?) and analytical method for measuring degradation (substance specific measurement = primary degradation or 14CO_2 evolution?) are not sufficiently clear as is the differentiation of the validity criteria for 20° and 12°C . Different validity criteria for open versus closed test setup may be required.

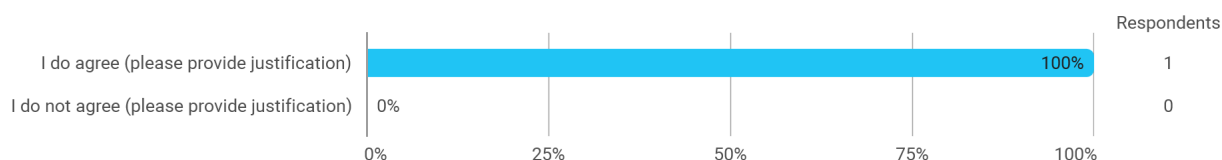
Rational: The OECD 309 guideline is suffering from some shortcomings, e.g. regarding pass criteria for test validity. Even more so in light of authorities' requests to conduct testing at 12°C . In order to enhance the applicability of the OECD 309 methodology with regard to the adapted temperature requirement and elucidation and assessment of relevant influencing factors to robustly assess the results of biodegradation simulation tests, LANXESS fully supports the current CEFIC LRI EC055 project (Assessing the Impact of Sample Collection on Microbial Population and Validity Criteria in the OECD 309 Surface Water Mineralisation Test).



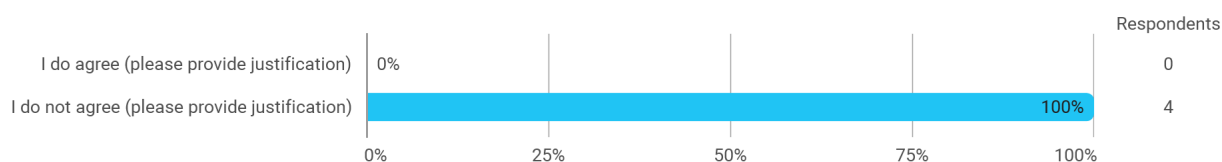
Agreement justifications

- Fully agree. Clear validity criteria are required for this TG. It has none at the moment, and that is a strong issue.
- Highly relevant, fully agree (comment submitted by myself...). There should be clarity for registrants and authorities regarding validity criteria under different test conditions. Furthermore it is crucial that the reference substance provides for a realistic representation of the microbial viability of the test system.

51. ANNEX 1 - DEFINITIONS AND UNITS, Degradation half time, DT50 (d):Maybe differentiate between DT50 for primary degradation and DT50 for mineralisation and report both



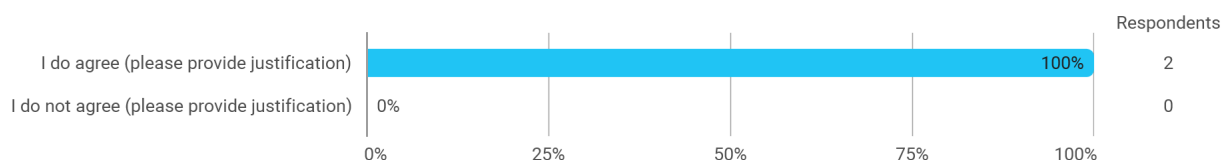
52. Increase the biomass to increase the likelihood of degraders being present. Consider potential for inclusion of a series of vessels in study design with increased biomass (alongside standard biomass) to reduce the biological lottery of not having competent degraders present.



Disagreement justifications

- ▶ The study purpose should be to test the degradation rate in a water sample that is environmentally relevant. Manipulation of the test system to increase degrader numbers moves away from this concept and adds confusion to the assessment of relevance of the outcome
- ▶ if at all only optional; manipulation of microbial community results in non-environmentally representative results
- ▶ Neither agree nor disagree: This might be useful but test would not be a simulation test in the strict sense.
- ▶ On the one hand, there is unclarity regarding the implementation. A precise specification on the implementation of systems with increased biomass should be given. On the other hand, test systems should be environmentally relevant (simulation test) and any manipulation of the test system to increase biomass moves away from this concept and reduces realism of the study.

53. Additional reference compounds in both screening (301 series, 302) and simulations (307,308,309) tests should be included. These additional reference compounds should allow to reflect degradability of more complex compounds ((bio)polymers).

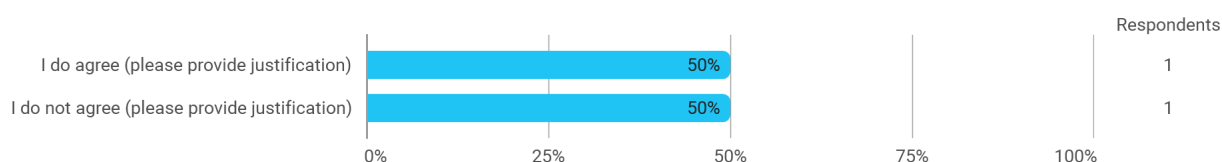


Agreement justifications

- ▶ We agree. The TG lacks relevant validity criteria.
- ▶ Agree: additional, reasonable reference substances for biodegradation but also for no occurring biodegradation (=persistence) should be included. In this regard, positive controls for detecting biodegradation should only follow this fate path which means a viable inoculum is required and should not have other properties prone to yield artefacts.

54. para 55 - The test should include more validity criteria. One of it could be that the test substance has stayed in the water. If not, the test is not valid. Also, the test should also give some range of acceptable degradation kinetics of the reference in function of the test temperature. Oxygen should not fall below 7 mg/L.

Rational: It is impossible to investigate degradability when the substance leaves the water phase by sorption or evaporation. In that case, the study should not be considered valid because it is not fit for purpose. A more reliable and clear validity criterion is required for the reference substance. The oxygen validity criterion is of utmost importance to ensure that the test accurately reflects an aerobic DT50, which is the purpose of the test.



Agreement justifications

- Fully agree. In particular, focus should be made on the oxygen levels in the water phase, in particular in close setup.

55. Strikethrough text:

Paragraph 25, first sentence: “or stirring”

Paragraph 25, line 4: “or use magnetic stirring” and “or stirring”

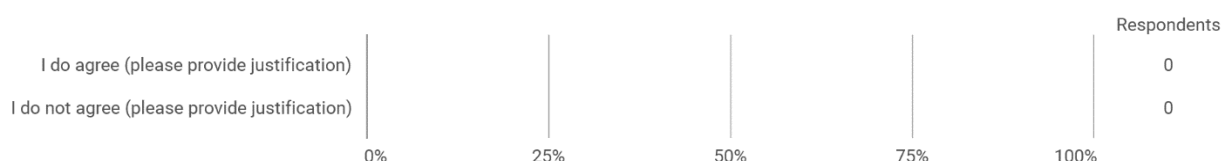
Paragraph 30, line 14: the second “substances”

Paragraph 31, fourth bullet point: “aniline or” instead insert “or [another suitable reference substance]”

Paragraph 34, lines 5-6: “position; in this case magnetic stirring and the use of magnetic bars coated with glass are recommended”

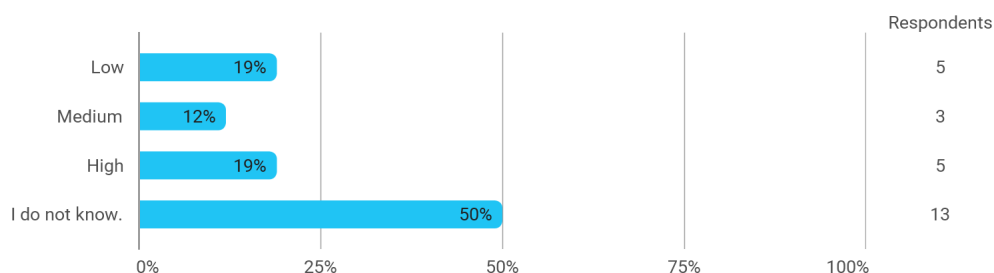
Paragraph 51, line 1: “aniline and” instead insert “

Paragraph 53, sixth bullet point following “Results”: “(if any)”

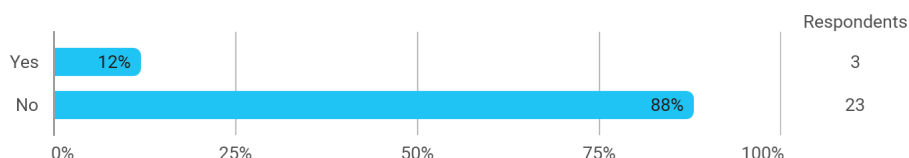


A.22 Test No. 111: Hydrolysis as a Function of pH

Relevance of revision

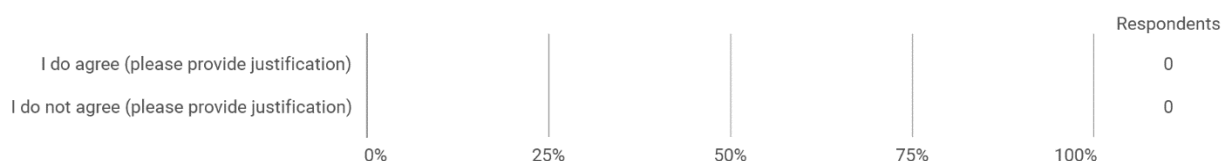


Interest in being part of an OECD Expert Group

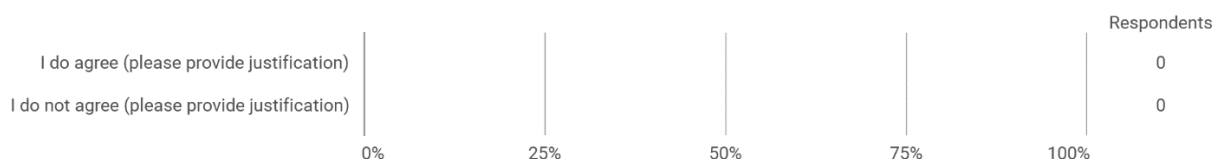


The following suggestions have been collected:

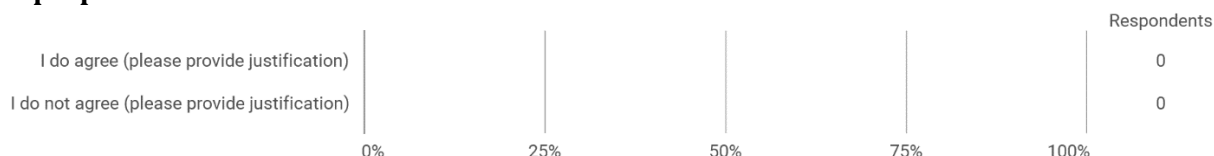
1. para 5- Incorporate an option to generate information on overall percent hydrolysis (instead of hydrolysis rate) for polymers. While it may be difficult to find precise analytical methods to quantify the hydrolysis rate of multi-component substances, this test can provide useful information on the general hydrolysis potential of certain polymers at various pHs. (e.g. by measuring changes in polymer size, properties)



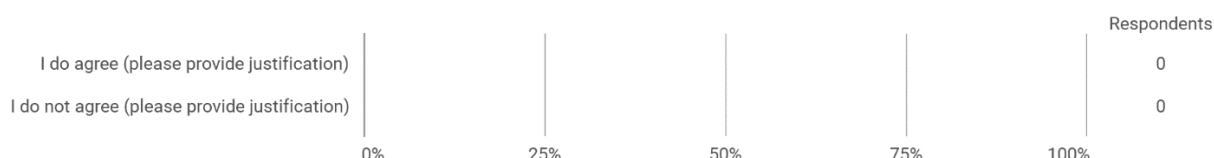
2. Add guidance for polymers



3. Refine preparation instructions

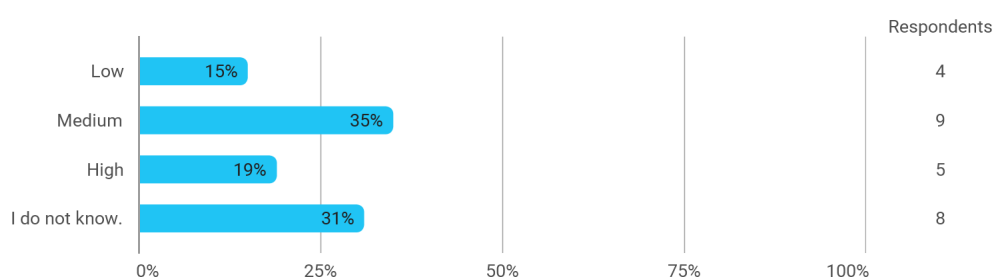


4. Refine instructions: The way of calculating and displaying the results can lead to wrong conclusions (e.g. relative area vs absolute area, normalization of the distribution curves); The TG states: "Any major hydrolysis products at least those representing > 10% of the applied dose should be identified by appropriate analytical methods." For oligomeric hydrolysis products, the concentration of a single chain length cannot be determined, how to decide which products need identification and quantification?

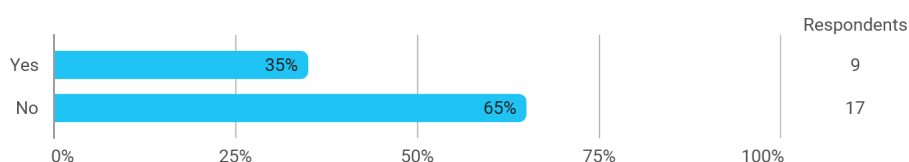


A.23 Test No. 316: Phototransformation of Chemicals in Water – Direct Photolysis

Relevance of revision



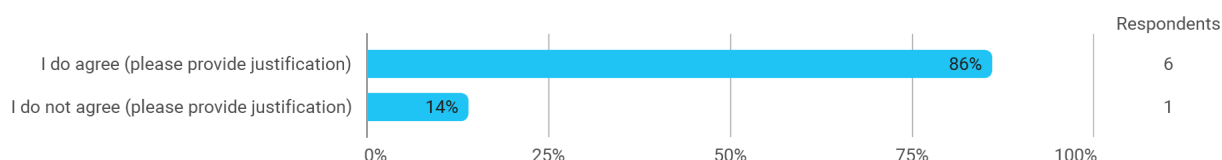
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

- 1. para 105 - It would be helpful if there is an agreed reference natural light intensity as the standard for conversion to summer sunlight. The natural sunlight intensity included in Annex 3 of the draft OECD soil photolysis guidance (2002) is commonly used in the conversion of data from aqueous photolysis studies as well, so it would be beneficial to include this in the updated guideline. The equation in Point 17 of the draft OECD soil photolysis guidance (2002) is also commonly used for conversion and should be included in OECD316. It would also be beneficial to finalise and issue the draft soil photolysis guideline**

Rational: Consistent conversion of experimental irradiation time to summer sunlight across users (for both DT50 and total test duration)



Agreement justifications

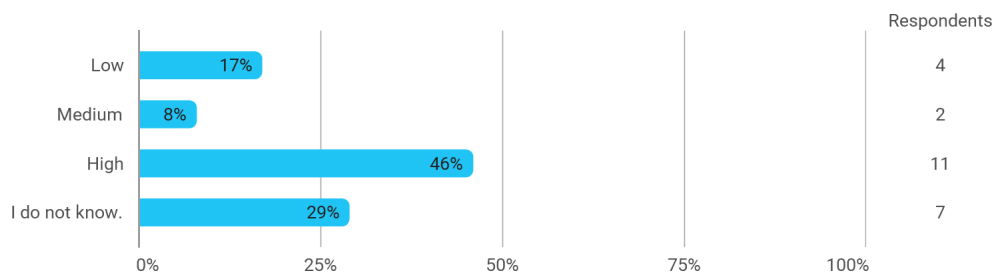
- ▶ Reference sunlight should be stated in more detail. Several more points are necessary to be questioned and discussed (e.g. use of software, how to get the irradiation for wavelength values with decimal places, setup of the test, etc.)
- ▶ The simple calculation in the draft OECD soil photolysis guideline has been used and accepted for many years despite it not appearing in any formalized guideline. This anomaly could easily be addressed by including this equation in OECD316. It is also very strange that the OECD soil photolysis was never finalised, meaning that this is the only study type in regulatory pesticide dossiers without an OECD guideline. My view is that the draft soil photolysis guideline should be finalized and issued
- ▶ This update would be helpful
- ▶ table with natural sunlight intensities should give a variety of global locations
- ▶ Comment agreed, but the required table of natural sunlight intensities should be extended by other global locations, e.g., Athens (Greece), Nanchang (PR China), Lianping (PR China), Edmonton (Canada), London (UK), Tokyo (Japan), Phoenix (USA) to have a set of standard values.
- ▶ Agreed; this would be very helpful.

Disagreement justifications

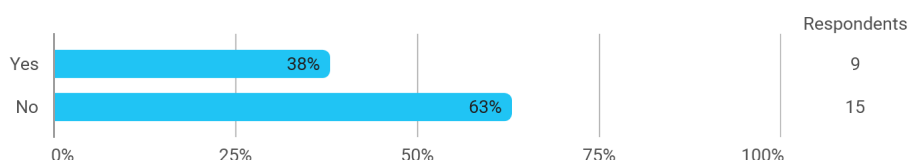
- ▶ There are already references to sun intensities for all seasons available in the photolysis guideline. Why summer-sun light? Why the connection to the soil guideline?

A.24 OECD Draft Soil Photolysis guideline (2002)

Relevance of revision

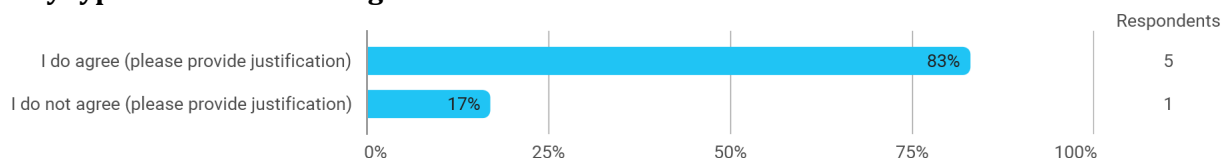


Interest in being part of an OECD Expert Group



The following suggestions have been collected:

- It would be helpful if the scope of this project could be expanded to include the potential for further development and finalisation of the 2002 OECD Draft Soil Photolysis guideline. There is no one aware of why this guideline has never been finalised and issued and it seems to be a significant anomaly that this is the only E Fate study type without an OECD guideline.**



Agreement justifications

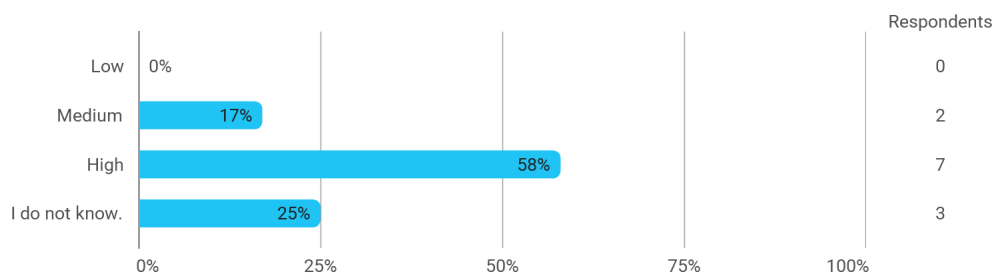
- Finalisation required
- This is a significant omission as this is the only study type that does not have an OECD guideline. It has never been clear why this draft was never issued and I feel that this should be rectified
- This guideline needs to be finalized, as it is a mandatory data requirement for registrations
- I support the suggestion. In addition, the scope of this study should be described in more detail, i.e., it should be emphasized that abiotic processes are the focus.
- Strongly agree; this guideline needs to be finalised after all this time.

Disagreement justifications

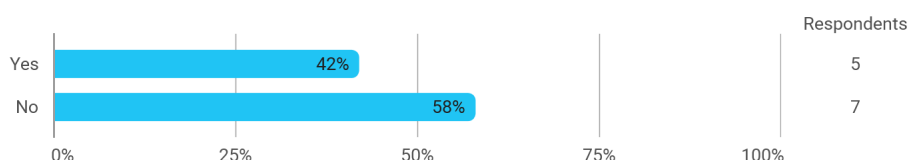
- This guideline cannot be finalized in its present form and version.

A.25 Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure

Relevance of revision



Interest in being part of an OECD Expert Group



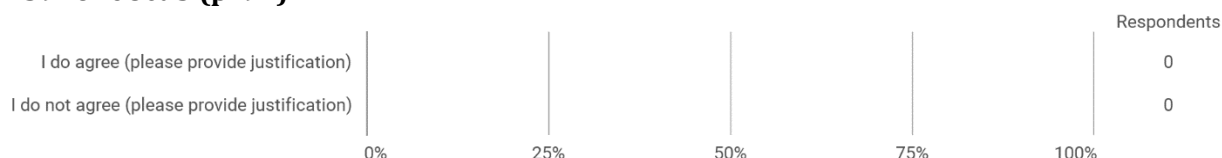
The following suggestions have been collected:

1. Growth dilution correction for kinetic BCF and BMF – improve the guidance for cases when the depuration rate is slower than the growth rate.

Rational: Growth dilution correction approaches didn't work for the MCCP case.

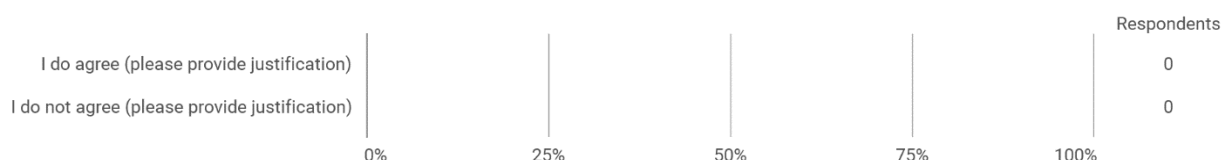
Please see more details in the report

published <https://echa.europa.eu/documents/10162/c4c72a94-44fb-bba0-ae17-875916108cd3> (p. 92).



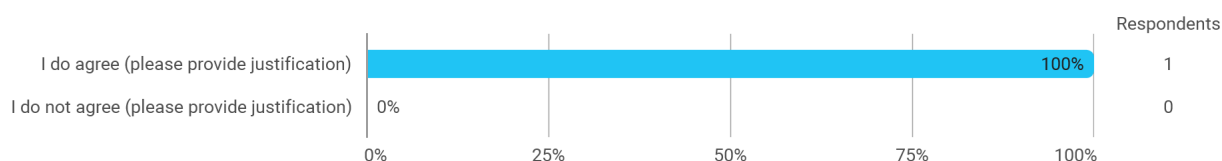
2. There are some bugs in the bcmfR R-package which should be corrected.

Rational: RIVM was contacted and is aware of the needs for updating the software package.



3. Concentration dependence of bioaccumulation is repeatedly discussed within the respective guideline. More clarification would be helpful for the cases where concentration dependence is indicated (what consequences for the data interpretation in such cases?).

Rational: In several sections of the Guideline, advice is given on how to conclude whether there is concentration dependence (e.g. 78). However, further consequences that follow from this conclusion are not reported.

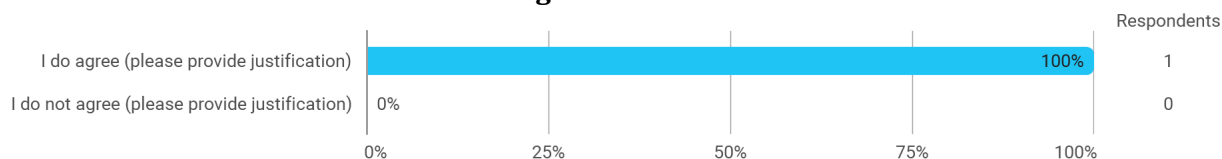


Agreement justifications

- Clear explanation missing

4. **para 16 - It is stated that a depuration phase is always necessary unless uptake of the substance during the uptake phase has been insignificant. In paragraph 38 it is stated that the test can be stopped if no significant uptake is shown after 28 days. However, the test guideline does not specify when uptake is considered insignificant. What is the cut-off value? Can this be clarified?**

Rational: As long as "insignificant" is not specified, the decision to stop the test and/or not to include a depuration phase is subject to interpretation. If the evaluator of a study reaches a different conclusion, the study may need to be repeated, which is not in line with the general objective to reduce animal testing. It would be very helpful to have a cut-off value for what is considered significant.

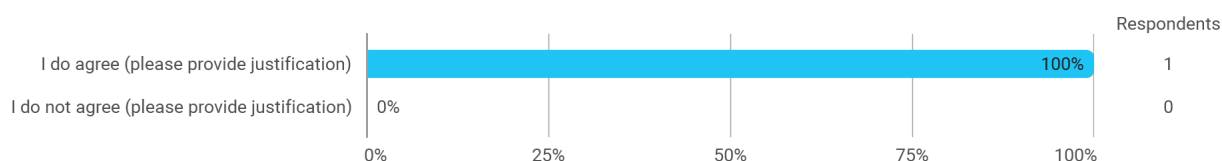


Agreement justifications

- Usually, BCF > 1000 correspond to a relevant bioaccumulation. Cut-off value proposal: 10% - > BCF 100.

5. **para 24 - The concentration of the test substance in the chambers is maintained within $\pm 20\%$ of the mean of the measured values during the uptake phase - this criterion may invalidate a study, e.g. when testing UVCBs where for some compounds deviations may be higher than 20%. Dietary uptake would be the choice here but authority requests prescribe the performance of the aqueous exposure. This criterion should be changed to where possible.**

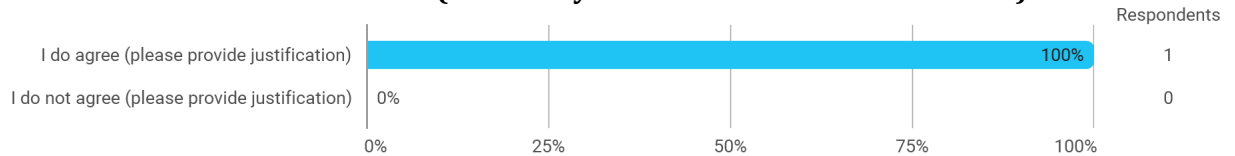
Rational: For UVCBs in the aqueous exposure no column elution / saturation columns are useful. Application should be done with solvents. The fate of single compounds of UVCBs in the water streams in an ongoing study cannot be influenced and may be affected by the presence of food, faeces, microbial growth, leading to a market decrease of concentration (outside the valid range). This is currently not considered in the validation criteria.



Agreement justifications

- With some substance not possible

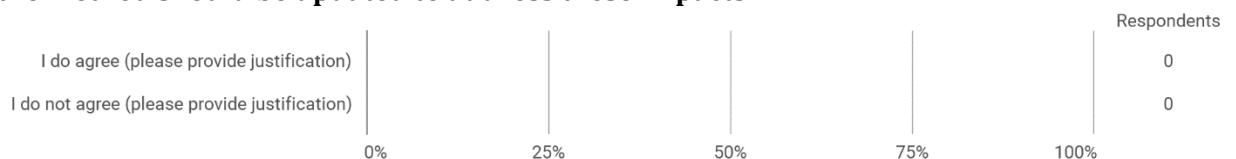
6. para 24 - Maintained at $\leq 20\%$ CV(variability of measured test concentration)



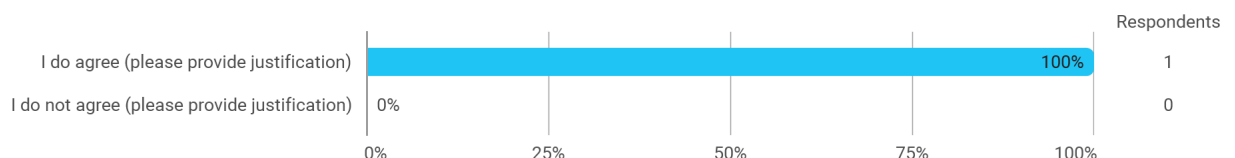
Agreement justifications

- compare OECD 231 (AMA)

7. **para 72 - A growth correction method was recently published by Dr. Frank Gobas. This peer-reviewed method should be incorporated into the test method to ensure fish growth does not impact the results of the test. This paragraph should also be updated to address the impacts of metabolism on bioaccumulation. The use of IVIVE and metabolism should be used in screening for bioaccumulation. Rational: Both growth and metabolism can significantly impact fish bioaccumulation and the method should be updated to address these impacts.**



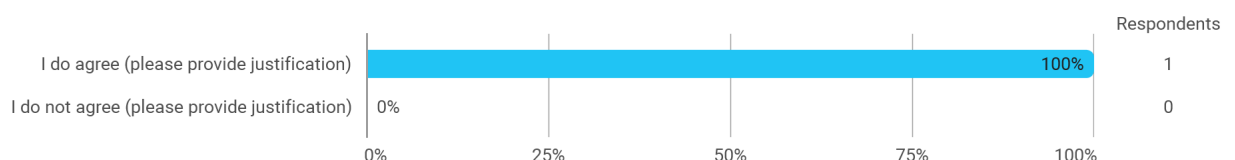
8. **para 99 - The dietary exposure pathway section should be updated to include the most recent guidance regarding criteria for identifying substances that biomagnify (BMF). Rational: The BMF criteria have evolved since the most recent version of the test guideline was published and should be updated accordingly**



Agreement justifications

- criteria for identifying substances that biomagnify

9. **Annex 3 - recommended temperature for rainbow trout: Lower temperature, change temperature to 12-16 °C**

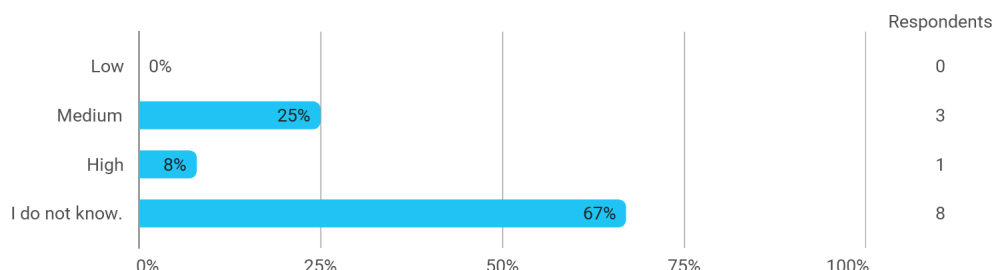


Agreement justifications

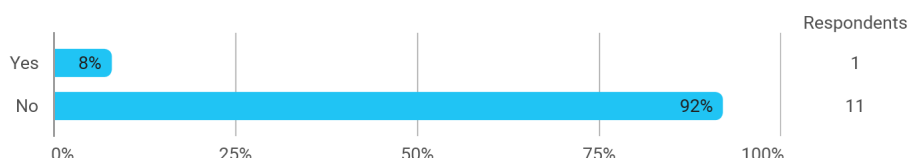
- trout from our supporter are used to lower temperature

A.26 Test No. 319B: Determination of in vitro intrinsic clearance using rainbow trout liver S9 sub-cellular fraction (RT-S9)

Relevance of revision



Interest in being part of an OECD Expert Group

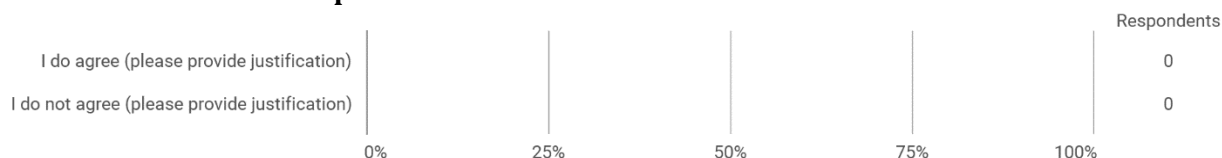


The following suggestions have been collected:

1. para 11 - Suggestion to include new optional addition of 'protease inhibitor PMSF' to increase the working lifetime of fish S9 reaction buffer, following this recent study: **Addition of Phenylmethylsulfonyl Fluoride Increases the Working Lifetime of the Trout Liver S9 Substrate Depletion Assay, Resulting in Improved Detection of Low Intrinsic Clearance Rates.** Nichols, J.W., Hoffman, A.D., Swintek, J.A., Droge, S.T.J., Fitzsimmons, P.N. *Environmental Toxicology and Chemistry*, 2021, 40(1), pp. 148-161

Rational: A potential disadvantage of the S9 assay is its relatively short working lifetime. This study evaluated that clearance rate just below the lowest detectable rate may still substantially influence BCF.

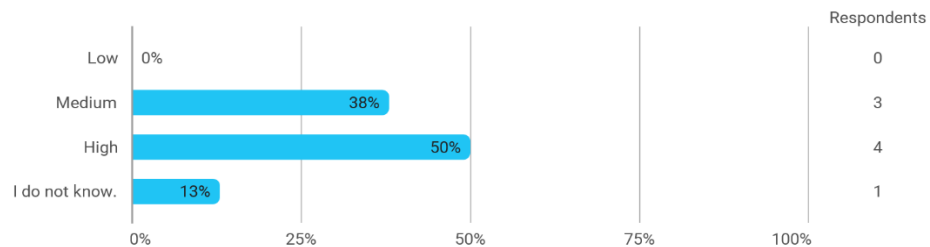
The protease inhibitor can extend the activity and lifetime of the S9 reaction buffer up to 6 h, and thereby sufficiently lower the detectable clearance rate to a level that could distinguish for most chemicals between relevant elimination rate to reduce the BCF or irrelevance of this process.



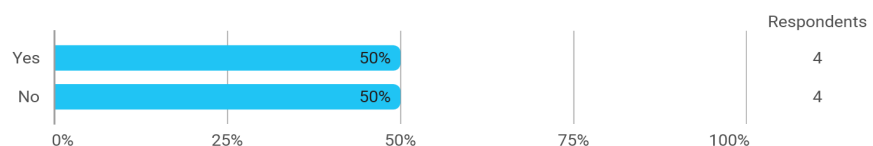
B Assessment feedback Block B

B.1 Test No. 205: Avian Dietary Toxicity Test

Relevance of revision



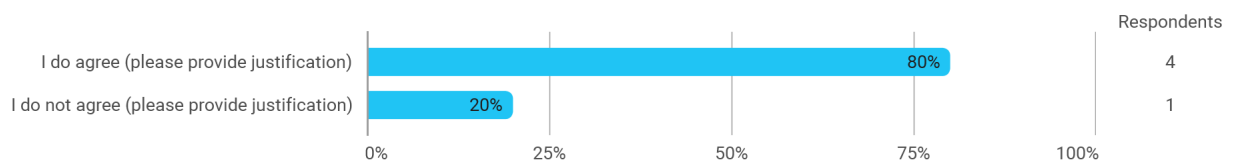
Interested to be part of an OECD Expert Group



The following suggestions have been collected:

1. General: Discontinue this guideline

Rational: Not required in EU and in phasing out by EPA; gives only very limited information of value for the RA; for modern compounds also of limited prediction value for dose ranges in the repro study (TG 206)



Agreement justification:

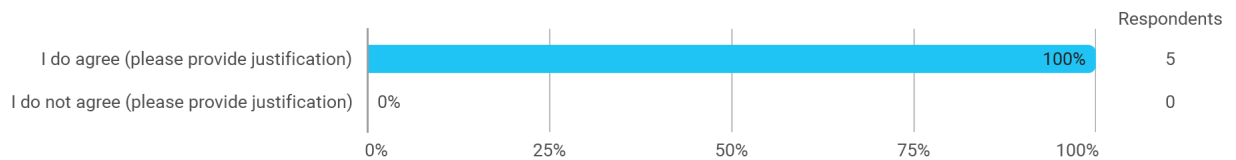
- No longer required in major regulatory areas
- This could be considered.
- Agree

Disagreement justification:

- While OECD 205 not routinely requested, in some instances multiple feedings e.g. an anticoagulant are more biologically relevant than a single dose

2. Table 1 - Revise age of birds, test species, and test conditions to comply with US guidelines

Rational: Avoid unnecessary rejection/duplication of tests



Agreement justification:

- It makes sense to align with US EPA's guideline that was updated in 2012
- if TG 205 is not discontinued, then the test conditions must be aligned to the other key test guideline for this test system

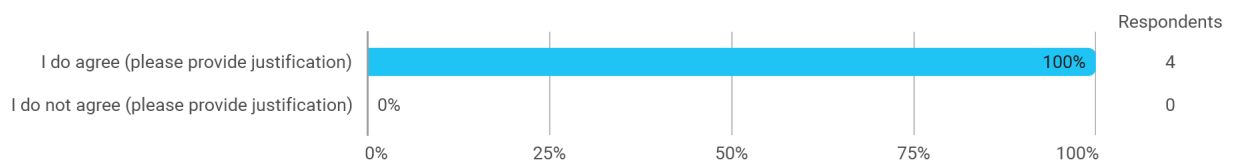
3. Adding a reference citing the US Environmental Protection Agency (EPA) Final Guidance for Waiving Sub-Acute Avian Dietary Tests for Pesticide Registration and Supporting Retrospective Analysis

(<https://www.epa.gov/sites/production/files/2020-02/documents/final-waiver-guidance-avian-sub-acute-dietary.pdf>) is requested.

Rational: The US EPA Final Guidance for Waiving Sub-Acute Avian Dietary states on page 3:

“... Several aspects of the avian dietary test that limits its utility in refined assessments:

- The study cannot provide a dose estimate for the effects endpoint because test organism consumption estimates are confounded by spillage, the lack of daily estimates of consumption, and mortalities occurring before study termination.
- The five-day exposure window is arbitrary, having more to do with laboratory expedience than any avian behavioral or toxicological factor.
- Toxicity is further confounded by the willingness of birds to consume food and the methodology cannot account for such behaviors as enhanced feeding rate during migration and the effect of assimilative energy differences between laboratory and field dietary matrices.
- Dietary concentrations are held constant during the study, limiting the use of food item degradation estimates in risk assessment.”

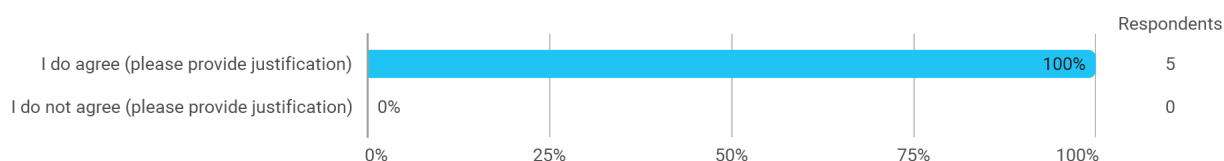


Agreement justification:

- This seems like a more reasonable approach than suggestion 1 leaving the test available for case by case

4. Reduce number of control groups (currently required in TG205: 2 control groups) - one control group is sufficient and nowadays common practice.

Rational: Comply with EFSA animal reduction act

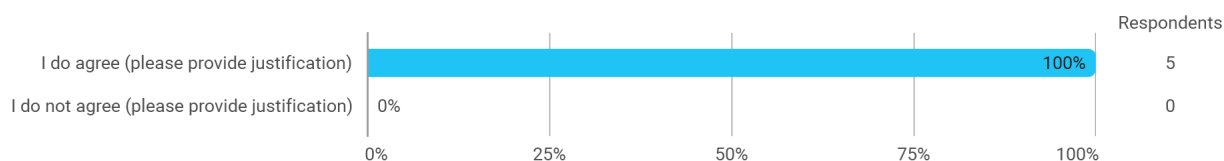


Agreement justification:

- Yes , one control group is all that's ever used nowadays
- experience has shown that one control group is sufficient

5. If the avian dietary test is to be conducted, the revision of guidance for 'Performance of the test' from two control groups to one control group is requested.

Rational: Harmonize to existing guidance which only requires one control group (negative control). See the US EPA OCSPP 850.2200 Avian Dietary Toxicity Test Guideline, section 5 on controls.

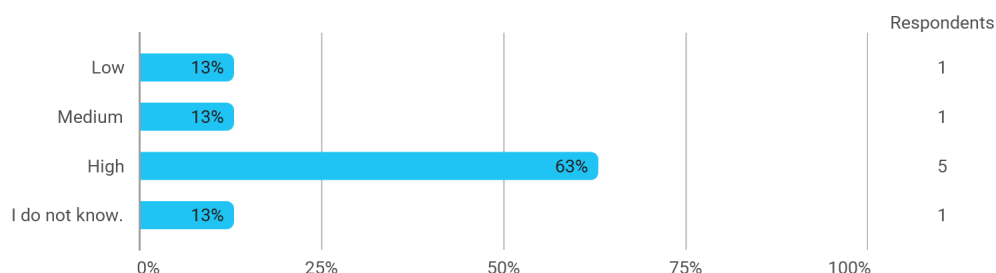


Agreement justification:

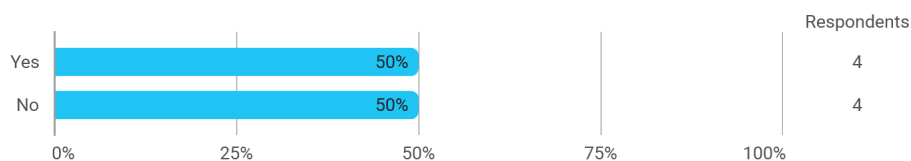
- Yes, as suggestion 4
- experience has shown that one control group is sufficient

B.2 Test No. 206: Avian Reproduction Test

Relevance of revision



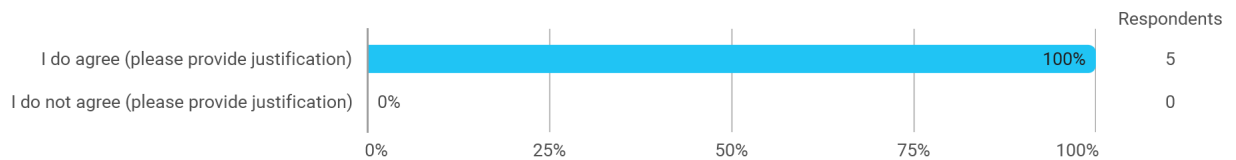
Interested to be part of an OECD Expert Group



The following suggestions have been collected:

- 1. All possible changes in the TG 206 must not lead to an invalidity of studies conducted under the current TG.**

Rational: Repeat of previously accepted/acceptable vertebrate studies should be minimized

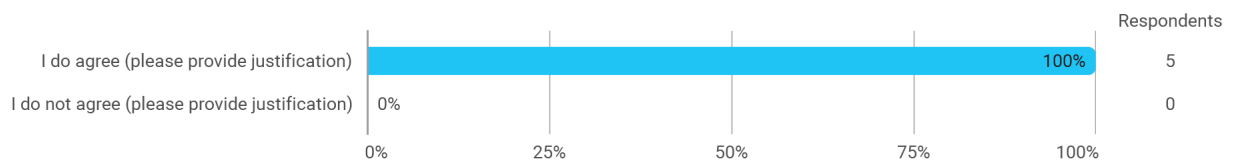


Agreement justification:

- Need to ensure that unnecessary vertebrate testing is avoided
- None of the proposed changes fundamentally question the relevance of the previous design, they just propose additions that will make conducting the study easier and options for more learnings with minimal changes

2. Revise the 0.34 mm required for mallard eggshell thickness to 0.32 mm to be in line with USEPA criteria.

Rational: Align with US EPA, 0.34 mm for eggshell thickness for mallards is also above the normal range encountered by mallards used at some laboratories

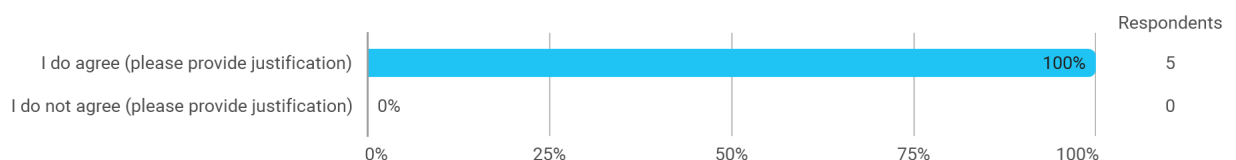


Agreement justification:

- Makes sense to align with US EPA, most of the studies are conducted in US labs anyway and the guidelines are harmonized. OCSPP 850.2300 was updated in 2012
- most TG 206 studies also aim to satisfy the EPA guideline, and 0.32 mm is normal and acceptable for Mallard eggshells
- We agree for a global alignment
- Global alignment.

3. Align test conditions with US guideline

Rational: Currently impossible to fully satisfy both guidelines in one study

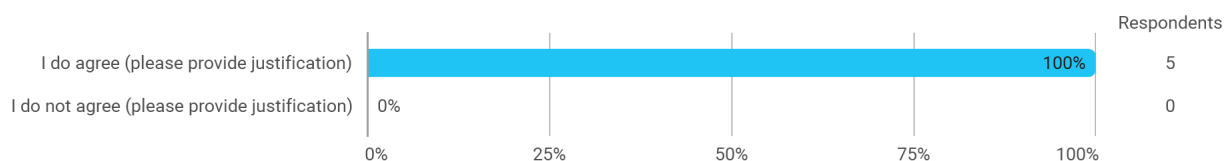


Agreement justification:

- I agree that it should be aligned with US EPAs guideline BUT only if there is not going to be a complete overhaul. I think a complete overhaul e.g. suggestion 13 or 33
- most TG 206 studies also aim to satisfy the EPA guideline,
- We agree for a global alignment
- Global alignment.

4. A harmonized agreement for translating ppm to mg/kg feed could be presented.

Rational: It would be quite simple to describe which steps to take in order to ensure that the same endpoint/value can reliably be derived

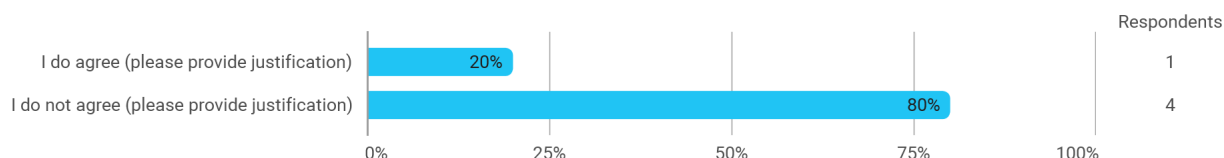


Agreement justification:

- I agree - a time weighted average would make most sense as birds are not weighed during egg laying yet they spend half the test at what is generally a much heavier weight than the first half of the test
- Assuming the proposal is for translating feed concentrations in ppm to achieved daily dietary doses in mg/kg bw/d: agree this is needed for EU risk assessment and it would probably be best to include the procedure recommended in the EFSA GD for Bird and Mammal Risk Assessment

5. Compared to other test systems like for mammals, a considerably reduced set of parameters is included. The development of a completely new guideline, in the best case with a 2-generation test design, should be considered.

Rational: The Guideline for the assessment of chronic and reprotox effects on birds is one of the oldest OECD guidelines and is therefore no longer considered to be the current state of scientific knowledge and technology.

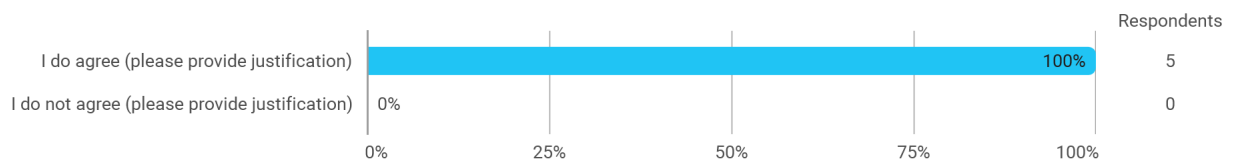


Disagreement justification:

- The avian 2-gen has not been adopted. This would use considerably more vertebrates and at present we do not know how some of the additional endpoints in the 2-gen would translate to apical endpoints
- Disagree with respect to developing a 2-generation study – this has been tried and not found to generate a more robust design. A 2 generation test would also lead to a very substantial increase in vertebrates needed for the test, without substantial benefit. As the future of testing will likely go towards less in-vivo vertebrate assays and more in-vitro and in-silico assessment, the TG should rather be amended towards options to include tailor-made investigations into aspects like time-to-effect, reversibility or ADME (eg, measuring residues in plasma or eggs)
- Disagree. Could be very difficult with avian species compared to laboratory bred rats and mice. Also, development time to reproductive phase is much longer in mallard/quail than rats/mice.
- Could be very difficult with avian species compared to laboratory bred rats and mice. Also, development time to reproductive phase is much longer in mallard/quail than rats/mice.

6. An inclusion of a recovery phase in the study design.

Rational: The exposure of a wild bird to a particular pesticides is in the natural environment mostly much shorter than in the avian reproduction study. This has to do with the use pattern of the pesticide and the degradation of the pesticides after application. This means, that the exposure in OECD 206 is normally a significant overestimation of the actual exposure that is expected under natural conditions. For example exposure to a seed treatment product is normally relatively short in the natural environment as this product will be applied only once per year. The possibility of the inclusion of a recovery phase could help to improve the actual risk assessment, especially for product for which a short exposure is only expected.

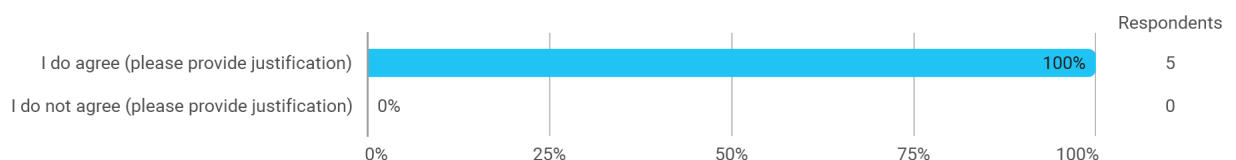


Agreement justification:

- ▶ Could be included as an option, already included in OCSP 850.2300, but guidance is vague (i.e., would need to provide info about sampling birds for necropsy, would egg collections and incubation continue). Must ensure that the extra information has sufficient statistical power, does not add vertebrates to the test while not taking away too much power from the rest of the test
- ▶ A general allowance of putting a test level in recovery phase can be useful in cases where it becomes obvious that a certain treatment level will clearly be no candidate for a NOAEL, because of clear and severe effect observations in the birds at this level. Then it would appear no good use of the test animals to simply continue until the end of the test. In such case, testing the reversibility of effects after cessation of treatment would add valuable mechanistic insights and allow to generate additional information and reuse the animals for a second purpose (6Rs).
- ▶ Agree. Could be very difficult with avian species compared to laboratory bred rats and mice.
- ▶ However, could be difficult with avian species compared to laboratory bred rats and mice.

7. Align validity criteria and 'normal values' with US guidelines.

Rational: Validity criteria and "normal values" should be revised to comply with historical control data.



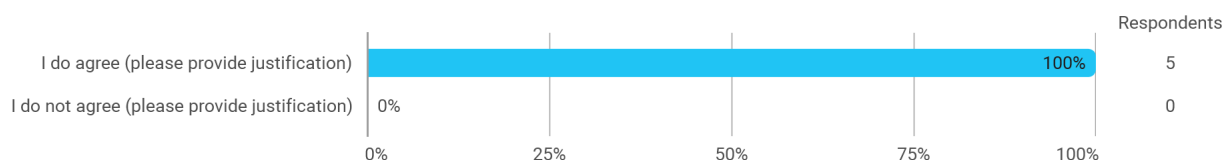
Agreement justification:

- ▶ I agree, although the preferred option is an overhaul of the current OECD 206 GD to something like suggestion 13 or 33
- ▶ most TG 206 studies also aim to satisfy the EPA guideline

- We agree for a global alignment
- Global alignment.

8. Remove feed consumption of hatchlings from list of observations.

Rational: Nearly never done (unreliable measurement), better captured in hatchling survival and 14-d survivor bodyweight which are much more robust assessment endpoints



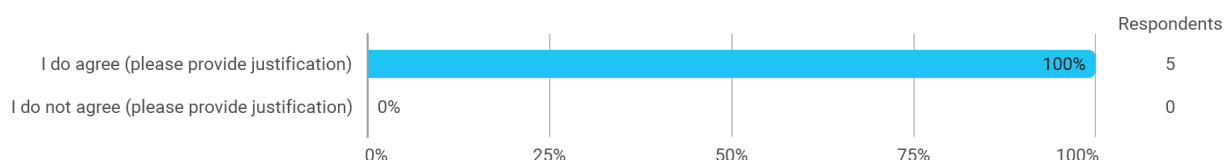
Agreement justification:

- difficult to measure, rarely done and the apical endpoint of growth will detect a significant effect on food consumption anyway
- Because of the technical difficulties to measure in a robust reliable way, feed consumption is nearly never performed, indicating that this determination is not requested by the reviewing authorities. Removing it from the list of required observations would reduce the risk of potential study repeat requests

9. Allow for inclusion of a recovery cohort which is only exposed for e.g., 15 weeks and then set on untreated diet to observe time to recovery.

Rational: Needed for PBTK and/or effect modelling which is expected to become state-of-the-art in the next years.

The recovery cohort could be added as a standard, but this would increase the number of birds even for studies with compounds with negligible bird toxicity (in which no recovery can be shown since there isn't an effect from which to recover). The preferred option is therefore to empower the study director to divide the treated groups where overt toxicity and large effects are already visible during the course of a study. Then, half of the birds at this affected treatment level would be retained on treatment, whilst the other half of the birds at this affected treatment would be used as the recovery group, where time course and magnitude of the recovery can be assessed in comparison to the other half of the treatment group which remains on treated diet.



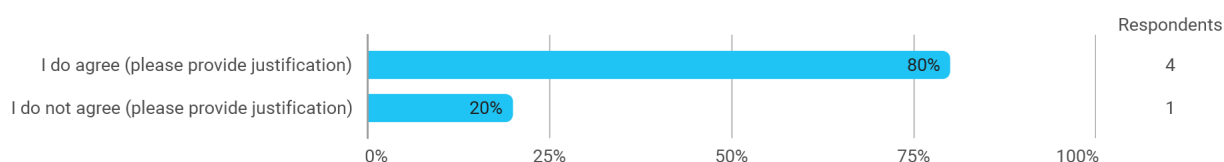
Agreement justification:

- Could be included - same response as suggestion 6
- Reversibility and the time course for recovery are important elements for developing and using PBK and/or effect modelling; technologies expected to become crucial in the next

future. In the long run, modelling may largely, possibly even fully, overcome the need for in-vivo testing. Recovery cohorts would be an important element for model development.

- Agree. Perhaps suggest the recovery period is a.i. specific based on phys-chem parameters, DT50 values, and crop residue data. Also, separate recovery cohort for seed treatment based on environmentally realistic seed treatment exposure periods dependent on crop type.
- Perhaps suggest the recovery period is a.i. specific based on phys-chem parameters, DT50 values, and crop residue data. Also, separate recovery cohort for seed treatment based on environmentally realistic seed treatment exposure periods dependent on crop type.

10. After “One or more species may be used for this test”, please add: “In some regions, the use of one species complies with regulatory requirements for testing of active substances of plant protection products.” Please add the references given below.
Rational: The number of animals used for testing should be reduced where possible.
Commission Regulation (EU) No 283/2013, section 8.1.1.1
India Central Insecticide Board and Registration Committee (CIBRC): Guidance Document on Toxicology for Registration of Chemical Pesticides in India.



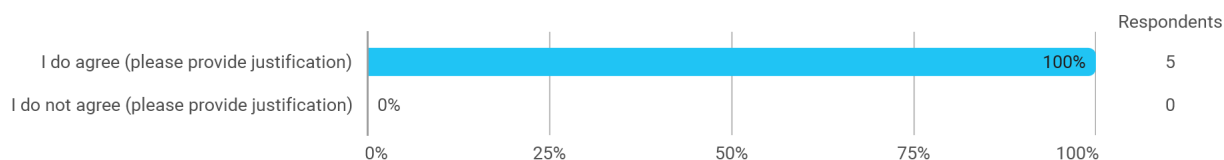
Agreement justification:

- Yes, makes sense to clarify and ensure unnecessary vertebrate testing is not conducted

Disagreement justification:

- I do not think that such detailed regulatory requirement information is needed in an OECD test guideline.

11. It should be clarified how the different endpoints can be compared and the dependency of some endpoints should be clearly described.
Rational: A discussion should be presented regarding the relative weight and interdependency of related endpoints. Often discussions arise from differences in the “overall” numbers and the comparison numbers (e.g., 14 d survivors vs 14 d survivors/viable embryos). It should be made clearer that comparison endpoints are important due to interdependency of the various endpoints.



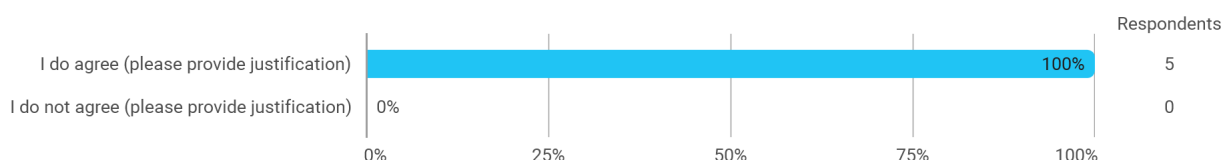
Agreement justification:

- Agree, clarification on need for concordance of endpoints would be valuable for data interpretation

- This comment is very vague, but may indicate a need to check for need of clarification by the experts in the actual revision working group

12. The number of replicates/individuals per replicate could be updated to allow effect concentrations (EC10/EC20) to be calculated more often. In the very least more clear information should be provided about the appropriate statistical analysis for the various endpoints from this test.

Rational: Avian reproduction tests are the basis of the pesticides risk assessment for birds in the EU, but the tests themselves very often results in difficult discussions because the statistical power of the different endpoints is low or the statistical analysis performed cannot calculate an effect level.



Agreement justification:

- As the EFSA 2021 draft update to the bird and mammal RA indicates that the preferred approach for analysis is lower limit of BMD 10, then it makes sense to adjust the test design to include more groups so it is better suited to generating such endpoints
- This would be a major and important change, requiring the agreement of global regulatory agencies
- Agree. But see ongoing CLE work on this.
- But see ongoing CLE work on this.

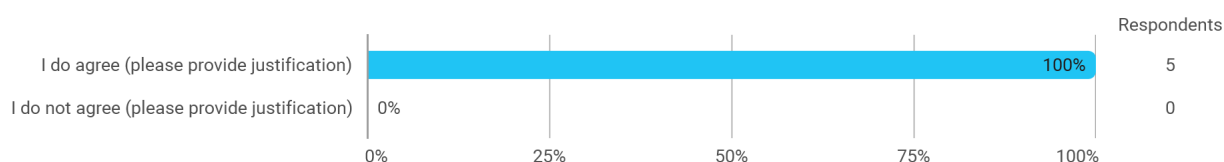
13. Suggestion to OECD 206 to change the current design from one that uses many reps (animals, typically 16 for mallard and 18 for Northern bobwhite) to one that is more similar to the expanded range-finding/preliminary test that is developed at Smithers (CRO). It appeared from the data, that it could be done in on the effects thresholds using more dose groups (at least 5 instead of 3), fewer pairs per group (5 instead of 16-18) and fewer number of egg sets (5 instead of 10). This test would need ring testing at the major avian testing labs (Eurofins, Easton [formerly EAG and prior to that Wildlife International] and Smithers). The acceptability criteria would need to be refined also.

Rational: This change would help reduce the number of animals used in tests from 3000-5000 to less than 1000 and still generate the same effects thresholds. It could also potentially be run without the range finding test (if include sufficient number of dose groups and span the doses broadly enough). Additionally, in my experience conducting these tests, the main reason for a repeated test is that there is no NOEC achieved. If you have more dose groups then you are less likely to have a rejected test because you could essentially use one dose group as a back stop to ensure that you test low enough to have a NOEC. If data in the original test are inconclusive, a repeat test could be conducted with an additional group plus a control (i.e., 2 pairs) and the two studies combined by looking at % inhibition relative to the respective concurrent controls. It may also be possible to use ECx, or Benchmark dose approaches for this

test design, by having more dose groups (although that would need ring testing). I have run many avian reproduction tests as study director and am now working in industry. I know first hand the number of animals that are used in the current OECD 206 test design and it really struck me with each test conducted that the guideline really needed to be tweaked to a minimized design with potential staged dosing to limit the number of animals used. I realize that there is appetite to replace vertebrate testing altogether, however I believe we are years away from that, and they may never be able to fully capture the complexity of the whole organism testing, Therefore, I strongly believe that we should invest the time and effort now to refine the existing OECD 206 test guideline to reduce the number of animals. I have presented on this topic at SETAC NA in 2021, see presentation title below and would be more than happy to work with anybody who shared my desire to modify this test guideline to reduce animal use.

Setac presentation

3.02.05 - Avian Reproductive Toxicity Tests - Could We Do More With Less? Thomas Bean¹, Tiffany Carro², Kelly Stanfield³ and John Schwalbe³, (1)Smithers ERS, LLC, United States, (2)FMC Agricultural Solutions, (3)Smithers

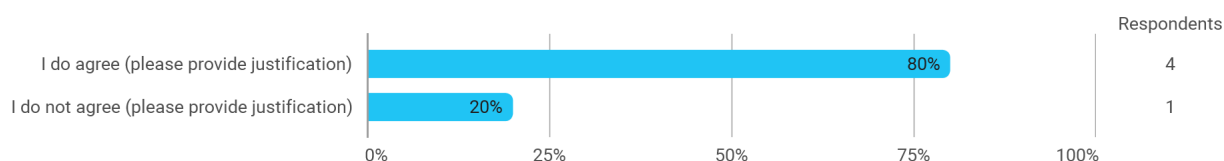


Agreement justification:

- ▶ This overhaul of the test guideline would enable a significant reduction in the number of vertebrates tested
- ▶ This would be a major and important change, requiring the agreement of global regulatory agencies
- ▶ Agree. Needs quantification via statistical modelling. Suggest consult statistical experts in working groups addressing these issues (e.g., CLE group proposing statistical approaches for avian repro data. See SETAC presentations and manuscript in review from the CLE M7 team).
- ▶ Needs quantification via statistical modelling. Suggest consult statistical experts in working groups addressing these issues (e.g., CLE group proposing statistical approaches for avian repro data. See SETAC presentations and manuscript in review from the CLE M7 team).

14. Remove „Avian dietary LC50 study (see Test Guideline 205).

Rational: LC50 test will no longer be always available (not required in EU and in phasing out by EPA)



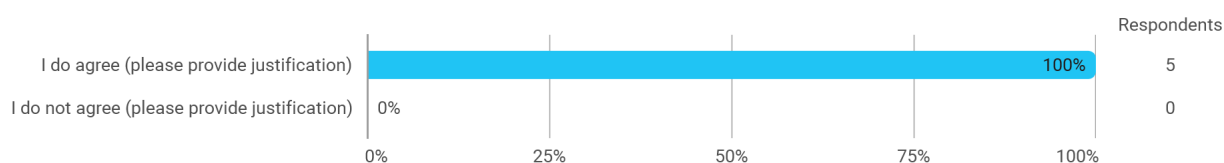
Agreement justification:

- ▶ Agree to remove cross reference to LC50 test (no longer routinely conducted) so not of great utility
- ▶ OECD TG 205 is proposed to be discontinued; if this will be implemented then the study type should also be removed in OECD TG 206 (or add a disclaimer “if available”).
- ▶ Agree.

Disagreement justification:

- ▶ Disagree. Suggest including potential for waiver following EPA guidance (see above).

15. Table 1 - Revise the recommended ages of the birds at test initiation in order to prevent egg laying prior to week 10. Based on data collected at Smithers, this should be 16-20 weeks for mallard and 16 to 30 weeks for bobwhite
Rational: In order to prevent eggs being laid prior to the 10th week of exposure

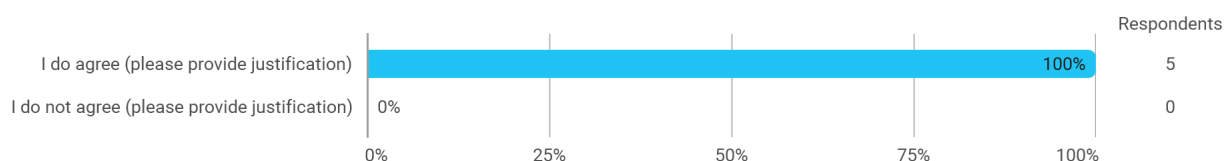


Agreement justification:

- ▶ I agree, standardization of the ages of birds would likely help to tighten inter-study control variability and also ensure that premature egg laying doesn't occur
- ▶ Agree: experience has shown the risk of early egg-laying is reduced with younger birds. Check with other labs than Smithers whether these proposed age ranges match their experiences.
- ▶ Agree. Consider suggestion from Smithers to be valid based on experience.
- ▶ Consider suggestion from Smithers to be valid based on their years of experience.

16. Table 2- Revise the humidity ranges during incubation to be more in line with modern day incubators. Consult with CROs performing the test to determine their ranges, based off experience, 48-60% for bobwhite and 50-60% for mallard is more reflective of what is used.

Rational: The guideline currently lists the humidity for incubation that is too high and not reflective of current practices using modern incubators. At 75% humidity (as quoted), mallards would not hatch, the eggshell would not dehydrate by enough during the incubation period to be able to hatch.

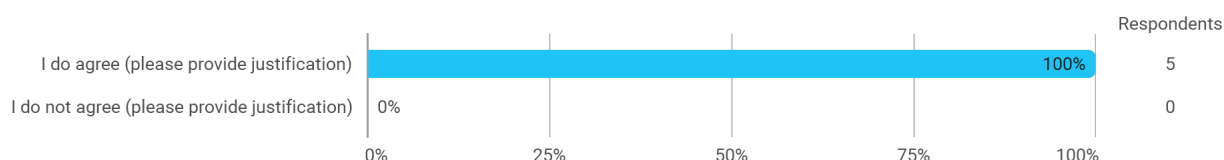


Agreement justification:

- Agree, existing ranges are not in line with the conditions used in modern incubators
- Agree, an update is desirable

17. Revise language on test substance concentration setting to remove ‘the highest concentrations should approximate ½ of the LC10’.

Rational: Since avian dietary studies aren’t required in the EU, the LC10 may not be available and/or range finding may suggest more suitable test concentrations

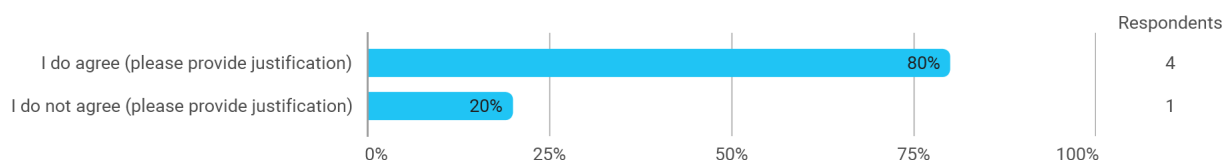


Agreement justification:

- As suggestion 14
- OECD TG 205 is proposed to be discontinued; if this will be implemented then the study type should also be removed in OECD TG 206 (or add a disclaimer “if available”). Suggest also to revise the reference of the ½ LC10 setting the top dose level – not achieved with most modern chemicals, rather add the reference to the NOAEL from the short-term dietary study (if available)
- Agree. ACR’s can be highly variable among active ingredients.
- ACR’s can be highly variable among a.i.’s.

18. Revise language saying ‘If a substance is stable only to the extent that diets would need to be renewed daily, the test may be inappropriate’.

Rational: This study is currently required for plant protection products in the EU, regardless of their stability in avian diet. If this study is not appropriate, what is the alternative study?



Agreement justification:

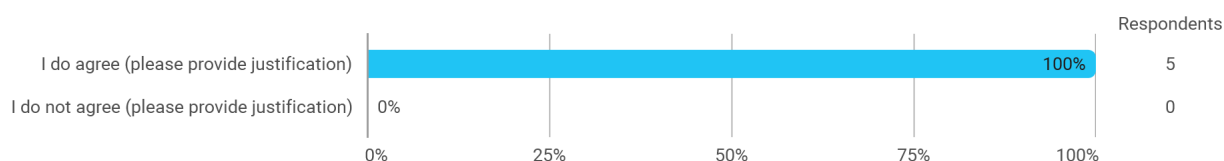
- I agree that this is a sensible recommendation as it is very difficult to change food out every day for 22+ weeks. However, an alternative option would be to permit TWA of dietary concentrations for unstable compounds with food changes every 48 hours

Disagreement justification:

- Very unclear which alternative is proposed

19. Revise day mallard eggs should be transferred to a hatcher to align with EPA guideline.

Rational: Currently impossible to satisfy both guidelines with 1 study

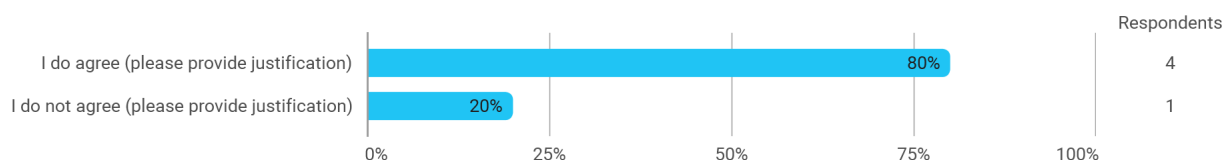


Agreement justification:

- Agree to align with OCSPP 850.2300
- most TG 206 studies also aim to satisfy the EPA guideline
- Agree for Global alignment
- Global alignment.

20. Reproductive history. Are birds included in the study first-time producers or have they produced eggs before the study began?

Rational: The responses or endpoint to be analyzed lack specificity, which leads to different measurements being reported by different testing facilities. This complicates inter-laboratory comparisons, compilation of useful historical control data, and increases the uncertainty of risk assessment



Agreement justification:

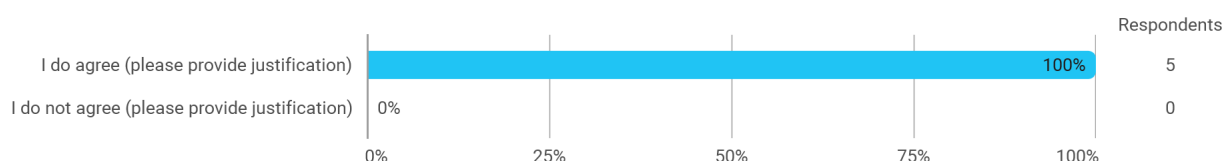
- Clarify that the intention is to use first-time producers
- Agree. It seems this should be noted.
- It seems this should be noted.

Disagreement justification:

- Unsure what this suggestion is. For mallard and bobwhite the current recommendation is for first time breeders, while experienced breeders are requested for Japanese quail

21. Food consumption. Is this measured on a per cage basis or per bird? Does it take account of food wasted?

Rational: The responses or endpoint to be analyzed lack specificity, which leads to different measurements being reported by different testing facilities. This complicates inter-laboratory comparisons, compilation of useful historical control data, and increases the uncertainty of risk assessment

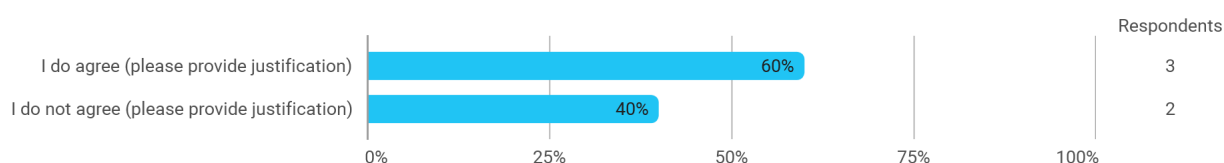


Agreement justification:

- ▶ Food consumption should be measured on a per pair basis and then divided by 2 to get a per bird basis. It is essential that wasted food is either recovered (quail) or estimated (mallard) otherwise the dose will be overestimated by 50% as revealed by a comparison between HCD at Eurofins where they do not recapture spilled feed in mallard studies (~150g/bird/d) vs Smithers (~100 g/bird/d) where spilled feed is estimated
- ▶ Clarify it is typically determined per cage and that efforts to minimize wastage should be undertaken (eg appropriate feeders). However, some food wastage will occur and is nearly impossible to quantify for both mallard and quail studies
- ▶ Agree. Important because daily dose calculations will be affected. It may not be feasible to separate male and female consumption if residing within the same cage. This can be adjusted with an adjustment factor that is established from a specific waste food experiment determining male and female feeding waste. For the second point, wasted food should be separated from mass of food not wasted and considered ingested.
- ▶ Important because daily dose calculations will be affected. It may not be feasible to separate male and female consumption if residing within the same cage. This can be adjusted with an adjustment factor that is established from a specific waste food experiment determining male and female feeding waste. For the second point, wasted food should be separated from mass of food not wasted and considered ingested.

22. The percent of cracked eggs can be based on the number of eggs per hen, the number of eggs set per hen, or the number of viable eggs per hen.

Rational: The responses or endpoint to be analyzed lack specificity, which leads to different measurements being reported by different testing facilities. This complicates inter-laboratory comparisons, compilation of useful historical control data, and increases the uncertainty of risk assessment



Agreement justification:

- ▶ Agree. Seems like one of the first two options would be more appropriate than the third option.
- ▶ Seems like one of the first two options would be more appropriate than the third option.

Disagreement justification:

- ▶ I do not understand this suggestion. As cracked eggs are not incubated, cracked eggs should be expressed relative to eggs laid
- ▶ Cracked eggs are not set, so the only reasonable relation is to eggs laid per hen

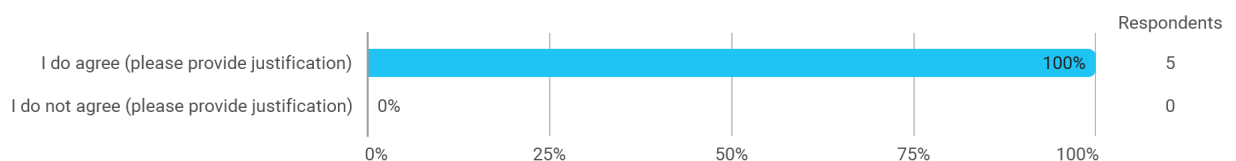
23. Reproductive parameters should be well defined for eggs and embryos.

Suggested endpoints:

Body weight parental birds

Body weight gain parental birds
Feed consumption parental birds
Eggs laid (ph)
Eggs not broken per eggs laid (ph)
Eggshell thickness
Viable embryos per eggs set (ph)
Live embryos per viable embryos (ph)
Hatchlings per live embryos (ph)
14-d survivors per hatchlings (ph)
Hatchling weight
14d survivor weight
14-d survivors per hen
ph: per hen alive at the end of the test

Rational: define logical, meaningful and consistent list of required assessment endpoints for reproductive parameter

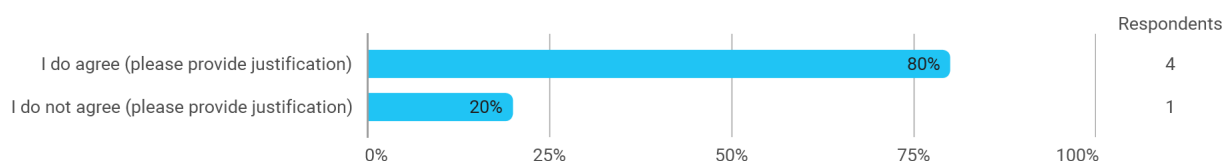


Agreement justification:

- ▶ General agreement that the endpoints should be aligned with US EPA. Suggest that the list of endpoints analyzed statistically just follows the revised list provided in CETIS support document 03
- ▶ The first 3 endpoints are not reproductive parameters but otherwise agree to the proposal
- ▶ Agree. Reproductive endpoints should be based on hens that are alive at the end of the study.
- ▶ Reproductive endpoints should be based on hens that are alive at the end of the study.

24. The use of model averaging rather than model selection to obtain a BMD estimate should be encouraged. EFSA guidance (2017) suggests basing the risk assessment on the lower 95% confidence bound (BMD10LB) on the estimated BMD10, whether on a single model or model average. Computer simulations indicate that for responses with high variability (CV=20 or higher), as many as half of studies will lead to model average BMD10LB values less than or indistinguishable from zero, making the BMD10LB useless for risk assessment. Use of the BMD10 point estimate or BMD20 would allow more responses to be analyzed through BMD methods (EFSA 2017. Update: use of the benchmark dose approach in risk assessment. EFSA Journal 2017;15(1):4658).

Rational: No guidance is provided other than a NOAEL/NOEC is to be determined for each response. While some flexibility in statistical methods is needed, general guidance is needed



Agreement justification:

- I agree to some extent, but if we are going down the BMD route, then also need to overhaul the guideline see suggestion 13 or 33
- Agree. Also see work by CLE subgroup publication and presentations on statistical approaches for avian reproduction study endpoints.
- Also see work by CLE subgroup publication and presentations on statistical approaches for avian reproduction study endpoints.

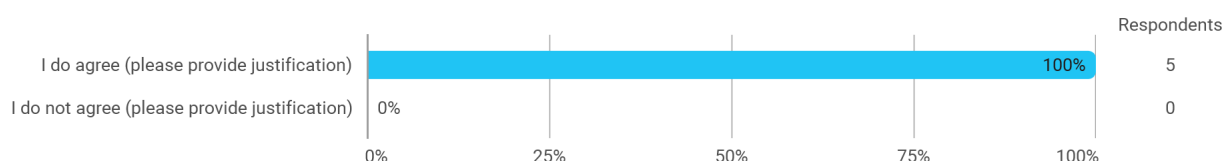
Disagreement justification:

- This comment seems misplaced

25. Caution is advised for any data manipulation prior to statistical analysis.

Transforming treatment data into percent or proportion change from control by subtracting the control mean and then dividing the difference by the control mean makes all observations correlated since the same random variable, control mean, appears in both numerator and denominator) of all such transformed data. Such a transform can be useful for data display but serious errors in NOAEL and BMD can result. See Green et al (2018) for details (Green JW, Springer TA, Holbech H 2018. Statistical Analysis of Ecotoxicity Studies. Wiley. ISBN: 978-1-119-48881-1).

Rational: No guidance is provided other than a NOAEL/NOEC is to be determined for each response. While some flexibility in statistical methods is needed, general guidance is needed

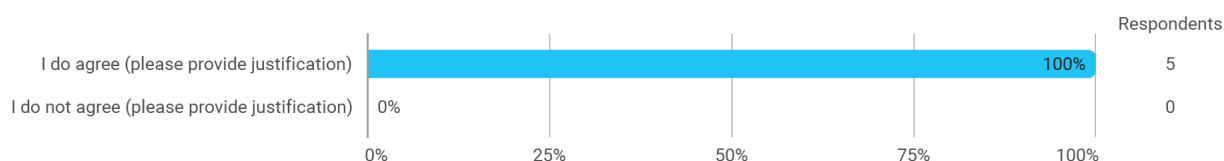


Agreement justification:

- I agree that caution is needed in manipulating/transforming data prior to analysis
- Agree, a clear recommendation which transformations make sense (and when and for what purpose) is important, but it must be harmonized with the text and the reality of the US guidelines because most studies aim to satisfy both guidelines

26. Define workflow and principle of the statistical tests, which should include outlier identification, and how to decide suitability of trend tests.

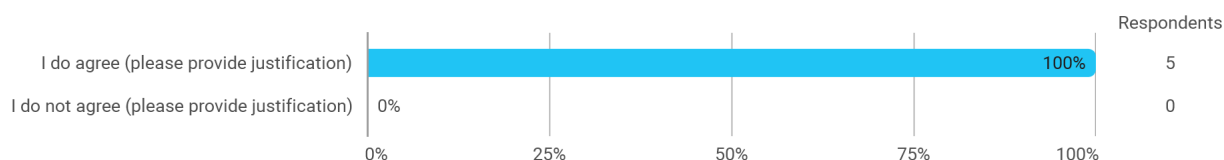
Rational: Statistical analysis needs to be clearly defined due to the high rejection rate of avian reproduction studies.



Agreement justification:

- A clearly defined decision tree for the comparisons tests would be welcome and reduce some of the uncertainty around interpretation of the test
- Agree, a clear recommendation which tests make sense (and when and for what purpose) is important, but it must be harmonized with the text and the reality of the US guidelines because most studies aim to satisfy both guidelines

27. Use of historical control data (HCD) should be an option for data evaluation. An appropriate HCD will identify an unusual control response that leads to spurious statistical significance. It can also distinguish between true effects and statistical artifacts arising from a mild trend in the concentration-response even though all treatment means lie within the HCD. To allow meaningful use of HCD, guidance on the construction and use of such a database is needed. EFSA (2017) recommends a 5-year window centred on the start date of the current study, modified, if necessary, to have a minimum of 20 studies at least approximately split between before and after the start date of the current study. Obviously, that applies only to evaluation of older studies. For a new study, there will only be historical data from current or earlier years. A mechanism should be identified to determine conditions under which data from a different testing facility can be part of the HCD for a given study. This could be a statistical demonstration on an individual endpoint basis of no statistically significant difference in results from the different labs whose data are used. Use of historical control data for avian reproduction tests has been evaluated by Valverde-Garcia et al (2018) (Valverde-Garcia P, Springer T, Vince Kramer V, Manousos Foudoulakis M, James R. Wheeler JR 2018. An avian reproduction study historical control database: A tool for data Interpretation. Regulatory Toxicology and Pharmacology 92: 295-302). Rational: No guidance is provided other than a NOAEL/NOEC is to be determined for each response. While some flexibility in statistical methods is needed, general guidance is needed

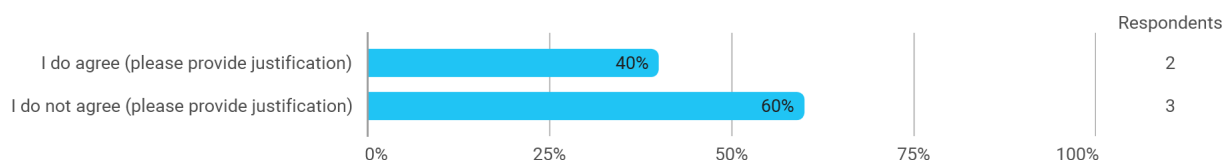


Agreement justification:

- Agree HCD is an important tool, also see suggestion 15 on how HCD might be tightened up going forwards.
- Agree that a general recommendation to consult HCD should be included, but would refrain from details because this is an evolving field, and any specific recommendation would likely be soon outdated

28. Where no suitable regression model can be found, the use of the MAXSD approach sometimes allows determination of the highest tested concentration at or below which there is significantly less than a 10% effect. The resulting MAXSD concentration is a BMD10LB.

Rational: No guidance is provided other than a NOAEL/NOEC is to be determined for each response. While some flexibility in statistical methods is needed, general guidance is needed



Agreement justification:

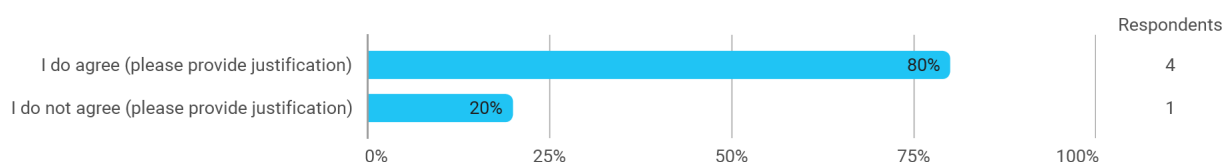
- Lower bound of BMD10 seems too conservative.

Disagreement justification:

- I do not understand this suggestion
- This comment seems misplaced
- Disagree. Lower bound of BMD10 seems conservative.

29. Data should be explored (visually or through formal tests) for consistency with normality and variance homogeneity if methods are used that require those data characteristics. Similarly, quantal and count data should be assessed for overdispersion and consistency with a Poisson or binomial distribution as appropriate.

Rational: No guidance is provided other than a NOAEL/NOEC is to be determined for each response. While some flexibility in statistical methods is needed, general guidance is needed



Agreement justification:

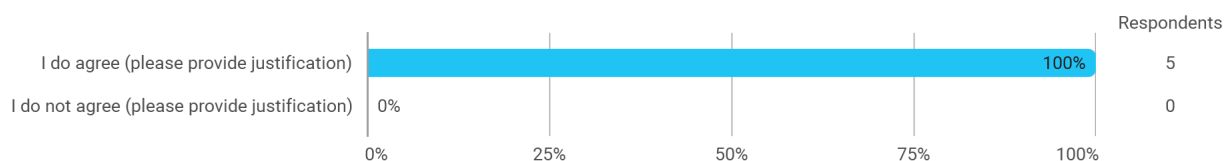
- Clear statistical guidance is welcome

Disagreement justification:

- Uncertain if an experimental test guideline must contain such general recommendation on how to treat specific types of data. Not a must

30. Both NOAEL and BMD methods have better statistical properties for some endpoints if generalized linear or non-linear models are used. Proportion data, such as eggs hatched per eggs laid can be bettered modelled as binomial conditioned on the number of eggs laid.

Rational: No guidance is provided other than a NOAEL/NOEC is to be determined for each response. While some flexibility in statistical methods is needed, general guidance is needed

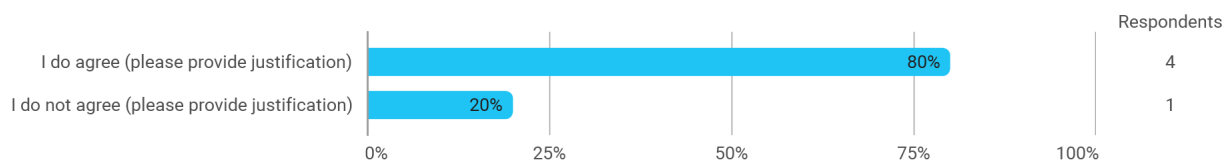


Agreement justification:

- ▶ OK, but would also need to increase the number of dose groups to better be able to have a dose response
- ▶ Agree to consider adding such proposals as long the general statistical recommendations are in reference to the actual test endpoints

31. Data should be explored (visually or through formal tests) for consistency with normality and variance homogeneity if methods are used that require those data characteristics. Similarly, quantal and count data should be assessed for overdispersion and consistency with a Poisson or binomial distribution as appropriate.

Rational: No guidance is provided other than a NOAEL/NOEC is to be determined for each response. While some flexibility in statistical methods is needed, general guidance is needed



Agreement justification:

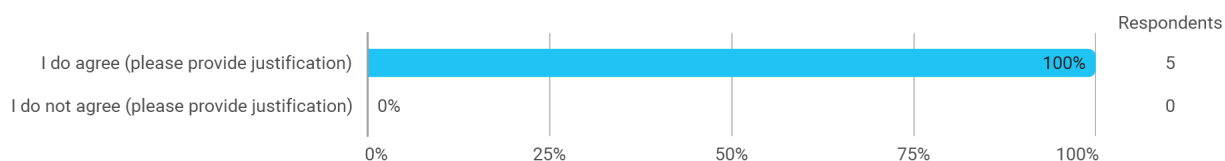
- ▶ Same as suggestion 29
- ▶ Agree. (Repeat of #29)

Disagreement justification:

- ▶ Uncertain if an experimental test guideline must contain such general recommendation on how to treat specific types of data. Not a must

32. Update age of adult birds and cage size according to current practice after > 35 years of guideline use in main avian testing facilities.

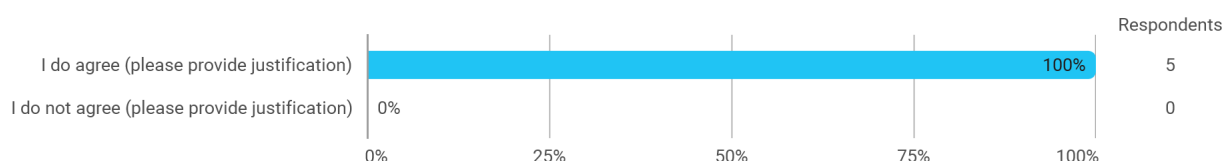
Rational: Comply with current practices for optimal reproductive performance.



Agreement justification:

- Agree that practices have changed since 1984

33. A minimum of 5 positive concentrations of the test substance in addition to a negative control are needed if the BMD/ECx approach is to be used for risk assessment. The current practice is to have 16 to 18 cages of paired birds per treatment group, including the control, and a total of three positive treatment groups for a total of 128-144 birds at study start. A design with 12 cages, 5 positive test concentrations, and a control would also require 144 birds. This would reduce the power of statistical test for a NOAEL determination by 12.5% compared to a design with 18 cage per treatment with only 3 positive treatment groups, but it would increase the likelihood of fitting a model for BMD estimation. To maintain the same power for NOAEL determination with 18 birds per treatment group would require 192-216 birds. Rational: No guidance is provided other than a NOAEL/NOEC. The experimental design should be modified to allow benchmark dose methodology (BMD) or regression analysis

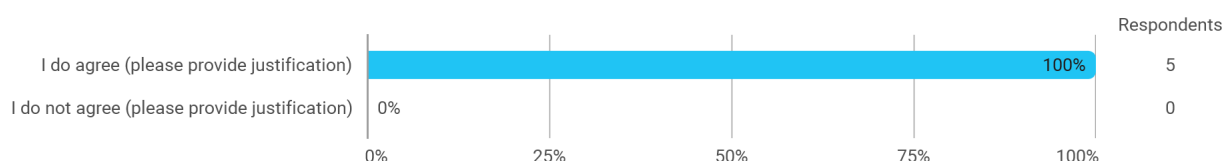


Agreement justification:

- An overhaul of the test design is really needed if we are to move towards ECx/ BMD type analyses. This approach would use 5 dose groups and the same number of animals
- Agree, such information is useful because BMD/ECx-designs may become important, and some guidance how to conduct these is necessary
- Agree, see CLE work
- I believe these power calculations were part of recent CLE efforts that have been presented and/or published.

34. Allow for a dose-response design that permits to reduce the number of replicates if the number of dose groups is increased.

Rational: Needed for EC10 or BMD analysis which is expected to become state-of-the-art in the next years



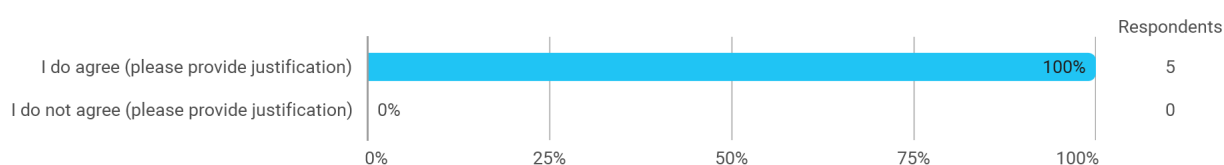
Agreement justification:

- Agree - with an overhaul of the test to increase number of dose groups this would be more likely to be successful

- Agree, such information is useful because BMD/ECx-designs may become important, and some guidance how to conduct these is necessary
- Agree (See #33 above).

35. Identification and assessment of effect of outliers on statistical analysis should be acknowledged. The Tukey outlier test for continuous responses is useful. An outlier is not a bad observation, just unusual, and warrants investigation. For example, 1 cracked egg out of 1 egg laid has a very different meaning from 35 cracked eggs out of 35 laid even though both yield a proportion of cracked eggs = 1. Either is likely to identified as an outlier.

Rational: No guidance is provided other than a NOAEL/NOEC is to be determined for each response. While some flexibility in statistical methods is needed, general guidance is needed

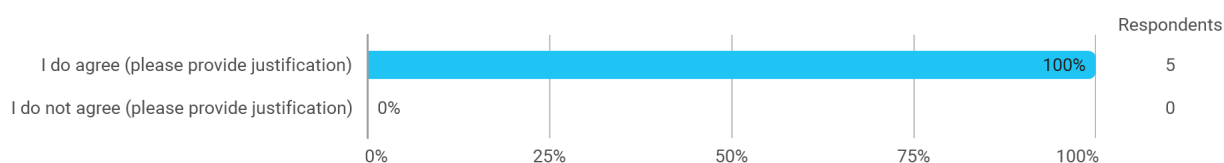


Agreement justification:

- Outliers are likely to occur in a test with terrestrial vertebrates, formal guidance on what constitutes an outlier and how to handle it would be welcome
- Agree, some general guidance on outlier considerations is needed

36. Allow for inclusion of small satellite groups (e.g., 3 males and 3 females per dose group) for taking blood samples and egg samples for residue analysis throughout the test (satellite groups in order to avoid interference with main test groups which provide the apical test endpoints).

Rational: Needed for PBTK and/or effect modelling which is expected to become state-of-the-art in the next years

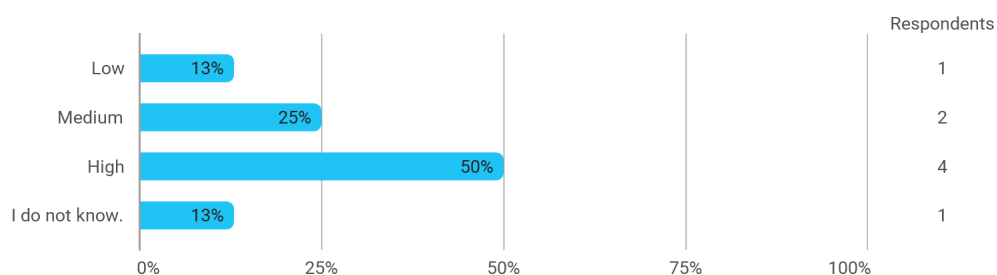


Agreement justification:

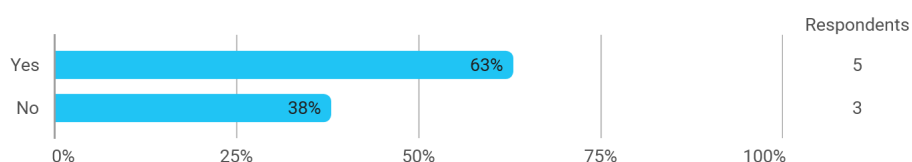
- Internal exposure and residues in eggs might help with interpretation of field data
- The importance of PBK modelling is likely to increase in near future, and such measurements are important for calibration / validation of models

B.3 Test No. 223: Avian Acute Oral Toxicity Test

Relevance of revision

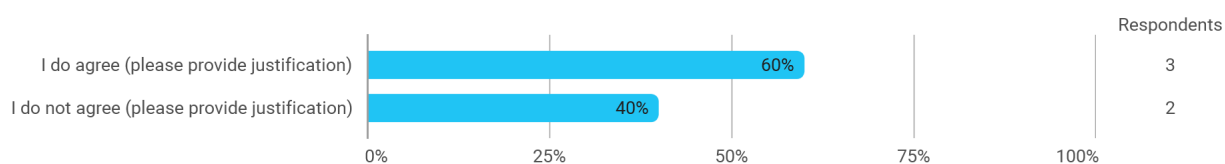


Interested to be part of an OECD Expert Group



The following suggestions have been collected:

- 1. Increase the number of birds in the limit study from n = 5 to n = 10 to be congruent with the EPA's OCSP 850.2100.**



Agreement justification:

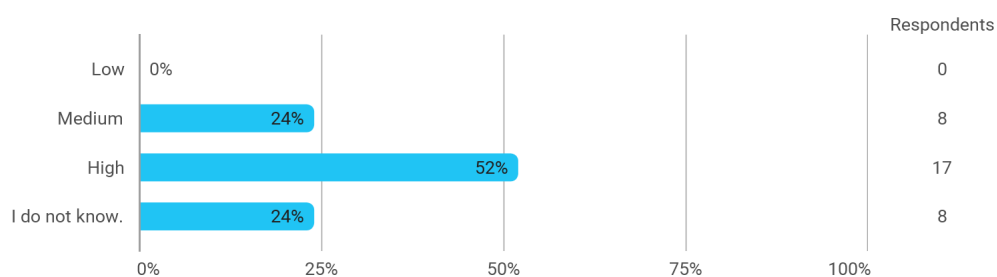
- Strongly agree!
- Strongly agree for global regulatory alignment.

Disagreement justification:

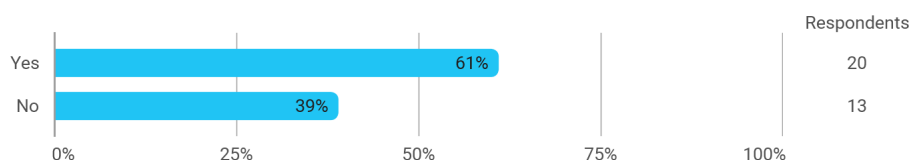
- I do not agree with increasing vertebrate testing
- EPA is accepting the OECD 223 study design therefore increasing the number of birds is not advisable

B.4 Test No. 213: Honeybees, Acute Oral Toxicity Test

Relevance of revision

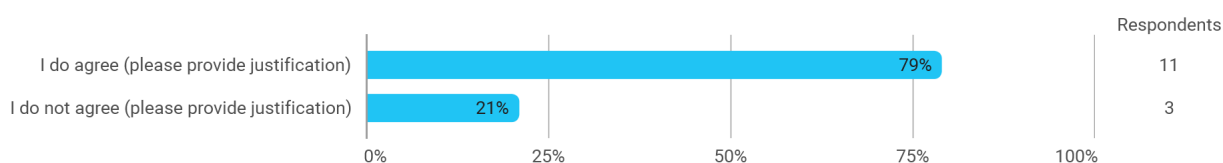


Interested to be part of an OECD Expert Group



The following suggestions have been collected:

- General:** Add a section on analytical verification of the dose. The text could be the same as in OECD GL 247, oral toxicity to the bumble bee, paragraphs 21 and 22.
Rational: The acute honeybee guidelines do not require analytical verification of the dose. All of the newer bee guidelines (OECD GL 237, 245, 246, 247 and GD 239 do include this requirement. Also in ecotoxicology tests with other organisms, analytical verification of the dose is common practice.



Agreement justification:

- ▶ correct
- ▶ analytical verification of exposure is a critical component of toxicity testing
- ▶ an aliquot of the lowest and highest dose sounds reasonable
- ▶ Agreed, analytical verification of the lowest and highest concentration of feeding solutions (+ stock where applicable) would provide verification of actual exposure levels.
- ▶ The dose verification should be carried out for OECD 213.
- ▶ Agree – we agree it would make sense to add analytical verification to this test guideline. However, concerns over the validity of all existing studies that do not have analytical verification if the guideline is changed to include it
- ▶ It is important to verify the intended dose was administered.

Disagreement justification:

- ▶ For contact this is not relevant at all - given that the lab technician knows how to prepare test solutions. for oral testing, the feeding duration is only for 4h (if at all, mostly bees consume the provided food much faster), it is not necessary to prove analytical stability. This would more be a control of the ability of the lab technician.
- ▶ We disagree, as this would unnecessarily complicate the performance of such a test. As the exposure is very short and the entire test solution is consumed, analytical verification should not be necessary for such a test system. In addition, in case clear precipitation of the test chemical is observed in the test solution by the qualified testing facility this is communicated and will result in lower test concentrations or a complete switch to another test item (i.e. formulated product instead of active ingredient).

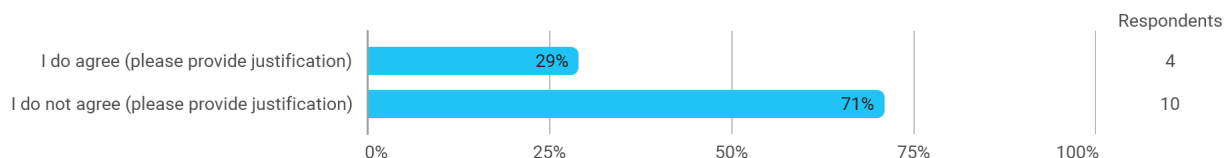
- The cited section is not suitable

2. para 5 - The test should have a duration of 96 h by default. This duration should be considered independently from the observed mortalities after 24 h 48 h.

Rational: According to Hesketh, H., Lahive, E., Horton, A. et al. Extending standard testing period in honeybees to predict lifespan impacts of pesticides and heavy metals using dynamic energy budget modelling. Sci Rep 6, 37655 (2016).

<https://doi.org/10.1038/srep37655>

it is recommended to extend the study duration to 96 h as standard to consider the different toxicokinetic properties of certain pesticides



Agreement justification:

- Standardizing acute toxicity tests to a common duration facilitates comparisons across chemicals.
- I agree only that the duration should be the same for all tests.

Disagreement justification:

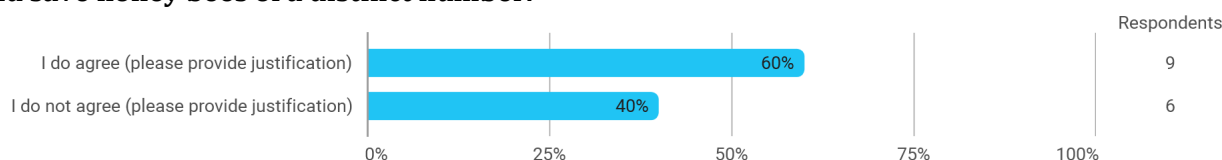
- Is not simply remains with a low mortality in the control group after 96 hours of ncaging
- This test simulates an acute exposure with a single application of a test item to differentiate between toxic and untoxic test items. As there is no 2nd exposure to the test item it is unlikely that mortality of bees starts later than 48h after test item administration. It is not necessary to prolong the test in general.
- duration of 96 h by default is not necessary, because when toxicity occurs the mortality will rise within the first 48 h
- This test assesses acute oral toxicity to honeybees and as such, the mortality during the first 2 days is of greatest concern. A prolonged observation period is only warranted in cases where mortality continues to increase between day 1 and day 2 in the test item group, while the control mortality remains below 10%. If the test duration was 96 hours by default, this information would not be of any benefit for the majority of compounds where mortality remains steady after 48 hours and would instead increase the cost/reduce the capacities available for this test design.
- Only increasing mortality > 10 %, the test should be prolonged.
- Disagree with a default of 96 hours, instead we would suggest clarifying the wording for when a study should be extended
- We disagree. As this is a test for acute oral toxicity to honey bees, immediate toxicity and mortality during the first 2 days are the main outcome. In cases where mortality continues to increase between day 1 and day 2 in the test item group, while the control mortality remains below 10%, the test duration is prolonged by default. Thus, setting the default to 96 h would

not be of any benefit for the majority of compounds where mortality remains steady after 48 hours.

- ▶ Disagree with a default of 96 hours, instead we would suggest clarifying the wording for when a study should be extended
- ▶ For test items with no/low toxicity a duration of 48 h is sufficient

3. para 6 - The toxic standard should lead to significant effects at a concentration of 0.20 µg a.i./bee.

Rational: Due to numerous studies, the effects and outcome of the studies with the toxic standard are well known by the laboratories that perform acute bee studies. Given the fact that animal welfare is getting more and more important, this change would save honey bees of a distinct number.



Agreement justification:

- ▶ Only one or two doses should be tested instead of generating an LD50. I would not add a specific dose but would more give a certain range within 50% mortality should be reached.
- ▶ sure, this should be fulfilled
- ▶ Agree - it would make sense to reduce toxic reference to 1 rate rather than multi rate – the exact rate would need to be chosen using historical reference item data and in agreement with the laboratories routinely conducting the studies
- ▶ one dose with effect would be sufficient to demonstrate the sensitivity of the system in all bee test. 'significant' has to be defined

Disagreement justification:

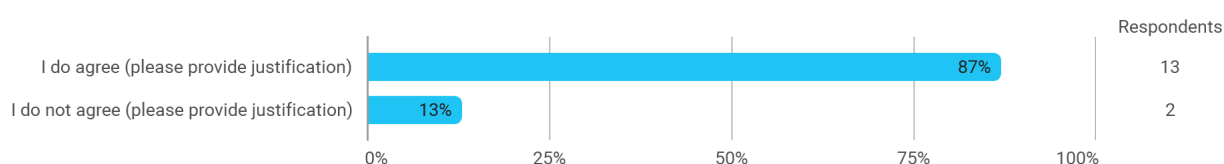
- ▶ Current wording is vague; statistically significant effect provides greater clarity, but it is not necessary to stipulate 0.20 µg ai/bee
- ▶ Disagreed, the historical range for the 24hr oral LD50 for the toxic reference dimethoate is between 0.1 - 0.35 µg a.i./bee. Either this range should be specified, or it should be clarified that a minimum of 50% effects is expected in the toxic reference group at the end of the test, independent of any particular dose.
- ▶ What means significant effect?
- ▶ We disagree with this comment. Toxicity values generated in different laboratories can be subject to certain level of variability. In standard laboratory toxicity trials for honeybees (OECD 213 & 214, 1998) a range for the toxic standard dimethoate is given (LD50 (24h) of 0.1 - 0.35 µg a.i./bee). This range was proposed after 12 consecutive years of data collection (Gough et al., 1994). Therefore, the target of this Guideline should be to allow the test laboratory to identify an appropriate test dose which will fulfil the intended validity criterium and not to prescribe a dose that may not be suitable. Gough, H. J., McIndoe, E. C., & Lewis, G. B. (1994). The use of dimethoate as a reference compound in laboratory acute

toxicity tests on honey bees (*Apis mellifera* L.) 1981–1992. Journal of Apicultural Research, 33, 119-125. doi:10.1080/00218839.1994.11100859

- This dose should not be a recommended dose as it is too high
- The LD50 should be the benchmark level and not a specific test rate.

4. para 6 - It could be clarified whether the first validity criteria is for the negative control, the solvent control or both.

Rational: As the criteria currently reads, one could consider that averaging all controls is a good practice. That could lead validation of a test where a zero per cent mortality was averaged with e.g. 17% mortality in the solvent control. However this test might not be considered as validity, since indicated adverse effects of the solvent used. It is advisable to clearly state in the TG that the role of using the solvent control is to demonstrate that the solvent did not influence the outcome of the test.



Agreement justification:

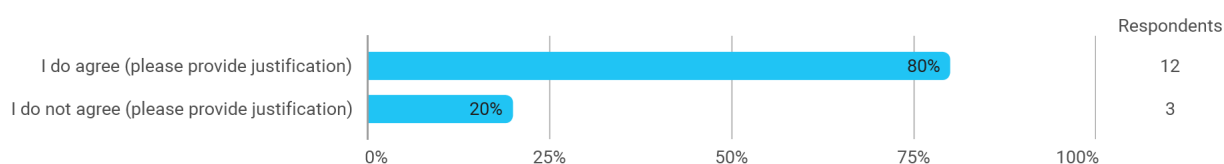
- can be clarified, but should be valid for all possible control groups
- Agreed, in the case where a solvent control is used the same validity criteria as for the control should be applied.
- In my opinion, both.
- We agree, no further justification needed.
- specify: control and solvent control
- also a mini guide should be needed on what to do with the solvent control

Disagreement justification:

- current wording implies that the criterion is for both negative and solvent control

5. para 6 - Par. 6: rewrite the sentence "the D50 of the toxic standard meets the specified range" with the following sentence: "mortality in the toxic reference substance group should be $\geq 50\%$ at the end of the test" (also for OECD TG 214).

Rational: To be in line with the new TG (e.g. 246, 247. 237) and to save some organisms



Agreement justification:

- Current wording is vague; the proposed modification provides greater clarity,

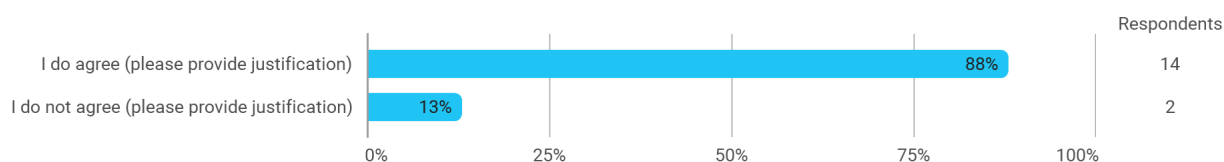
- ▶ Only one or two doses should be tested instead of generating an LD50. I would not add a specific dose but would more give a certain range within 50% mortality should be reached.
- ▶ valid, when just one concentration will be used for verification of the test system
- ▶ Agreed, the validity of the test is fulfilled where the toxic reference group shows significant adverse effects relative to the control.
- ▶ We agree, no further justification needed.
- ▶ One dose of the reference item is enough

Disagreement justification:

- ▶ I think "LD50-24 h is in the range of 0.10 - 0.35 µg a.i./bee" is better than only one dose rate.
- ▶ The LD50 is the important metric.

6. paras 6 and 16 - One or two test doses could be enough to demonstrate the sensitivity of the system in place of a full dose-response test for the toxic reference.

Rational: The aim of the toxic reference is to demonstrate the sensitivity of the system. To do that, a full series of doses to get a dose-response and derive a proper LD50 might not be necessary. Over the years, the method demonstrated a high level of repeatability. OECD No. 237 (and OECD GD 239) require only one test dose.



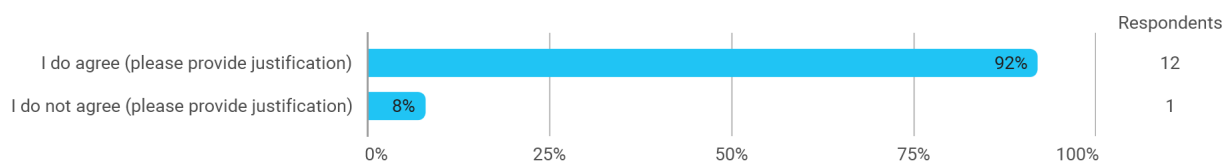
Agreement justification:

- ▶ Only one or two doses should be tested instead of generating an LD50. I would not add a specific dose but would more give a certain range within 50% mortality should be reached.
- ▶ totally agree, to save bees
- ▶ one dose resulting > 50% effect would be sufficient
- ▶ Agreed, most laboratories have a good understanding of where to expect a significant effect from the toxic reference item and in an effort to reduce the number of bees to be sacrificed for testing, the toxic reference could be restricted to a single dose, as is the case with all other laboratory tests with bees.
- ▶ We agree. However, each laboratory knows best which test dose can elicit the effects needed to fulfil the necessary validity criteria. Therefore, we advise against stipulating a pre-defined dose as toxicity values generated in different laboratories can be subject to a certain level of variability.
- ▶ One dose should be enough
- ▶ one dose resulting > 50% effect would be sufficient
- ▶ Only if a lab can show a consistent response over time

Disagreement justification:

- ▶ proposed wording is vague
- ▶ But how?

7. para 7 - They could be collected in the morning of use or on the day before test....



Agreement justification:

- ▶ It is not necessary to stipulate when during the preceding day bees should be collected.
- ▶ we already do it accordingly in the morning of use. Thus, I would have to guess if the transfer of the bees on the previous day has an influence. But I assume the oral test by food consumption at least
- ▶ Agreed, this wording provided for a bit more flexibility in the collection timing without impacting test performance.
- ▶ keep this sentence
- ▶ We agree. However, each laboratory knows best which approach works for them. Therefore, we advise to allow for flexibility and not be too prescriptive.
- ▶ Agree. However, each laboratory knows best which approach works for them. Therefore, we advise to allow for flexibility and not be too prescriptive.
- ▶ Both methods are suitable
- ▶ Only if it can be demonstrated that the level of handling stress is not an issue when the study is started.

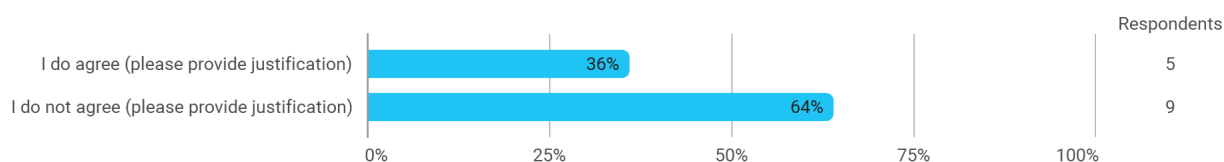
Disagreement justification:

- ▶ Collecting foragers is not time consuming and for the comfort of bees I would not recommend collecting them the day before.

8. para 7 - The wording in paragraph 7 (collection of bees) should be changed to: **Young bees (max. 2 days old) reared out from brood combs taken from queen-right colonies that have no symptoms of diseases and that have a known maintenance and physiological status history should be used for the test. No chemical substances (such as antibiotics, anti varroa treatments, etc.) should have been used in the hive for at least one month prior to the test. If one colony cannot provide the appropriate number of bees, comb(s) from several colonies may be used. In this case, it is ensured that the bees are equitably distributed across the treatments. Brood frames with capped cells that are expected to hatch on the same day can either be incubated in a climatic chamber or be kept without nurse bees in a worker excluder box within the hive until hatch. In the first case sufficient food supply should be ensured either by honey and pollen which is on the same brood comb or by an additional comb**

containing food. One day before the test starts, the bees can be collected from the combs and distributed into the test cages. Anesthetisation should be avoided during collection. Bees should be acclimated to test conditions for about one day (after a hatching period of one day). Bees are to be fed with sucrose solution ad libitum, but no additional feeding of pollen or water is necessary during the acclimation and test period. No starvation period is necessary before test start.

Rational: The required exact age of bees is not set in OECD 213 and 214 and it is simply stated, that the bees should be "young". To allow for a longer study duration and to simplify the fulfillment of the validity criteria, bees should be collected according to OECD 245. This however also leads to an adjustment of the test conditions, which should be identical to OECD 245 to account for the freshly hatched and acclimatized bees.



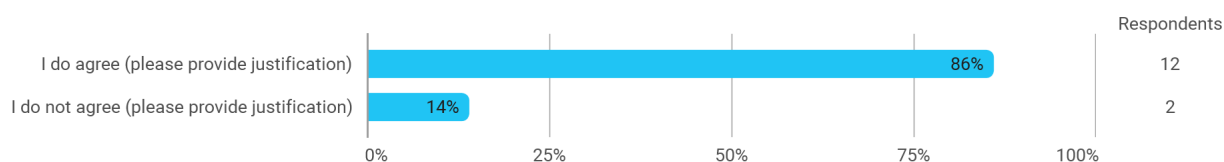
Agreement justification:

- Proposed wording better ensures consistency.
- Agree - however we would suggest this proposed wording should be reviewed by those at laboratories who routinely conduct the studies (if this comment did not originate from them).
- Some parts are in opposition to Suggestion 7 - acclimation time

Disagreement justification:

- I would like to know the justification for that proposed change
- absolutely disagree, because forager bees should be used and not "nurse bees"
- This is an acute test to simulate acute exposure of foragers bees in the field while or shortly after application of a test item. Young bees at the age of 2 days are constantly in the hive and have no acute exposure outside. Young bees are used in chronic testing as this is the realistic exposure in nature.
- the acute test is a test based on forager, not nurse bees. When young bees (nurse bees) will be collected a reduced food consumption occurs and the mortality can be with a high variability
- Disagreed, this wording is taken from the OECD GL 245 assessing chronic exposure on young bees, where age synchronization and standardization are critical to maintain control validity criteria. However, experience has shown that young bees take longer to get used to the feeding system (i.e. syringes) and may not consume the treated diet within the 3-4 hours (max. 6 hours) that constitute an acute exposure scenario. It is highly recommended to retain the wording of paragraph 7 to use bees from the hive and located on frames without brood (i.e. older than 2 days).
- That is for OECD 245 chronic oral toxicity test. Bees in an acute test should be a foraging bee.

- We do not agree. This wording is taken from the OECD TG 245, HONEY BEE (APIS MELLIFERA L.), CHRONIC ORAL TOXICITY TEST (10-DAY FEEDING), where age synchronization and standardization are critical to achieve the validity criteria, especially for the control group. However, experience has shown that young bees take longer to get used to the feeding system and may not consume the offered diet within the 3-4 hours (max. 6 hours) as required for an acute study. It is highly recommended to retain the wording of paragraph 7 to use bees from the hive and located on frames without brood (i.e. older than 2 days). Also, as honeybee foragers are expected to be the highest exposed caste it does not make sense to test the youngest bees which remain in the hive for the first 2-3 weeks. This is also relevant to consider for OECD TG 214 which simulates an exposure of a spray application.
 - Disagree - this wording is taken from the OECD TG 245, HONEY BEE (APIS MELLIFERA L.), CHRONIC ORAL TOXICITY TEST (10-DAY FEEDING), where age synchronization and standardization are critical to achieve the validity criteria, especially for the control group. However, experience has shown that young bees take longer to get used to the feeding system and may not consume the offered diet within the 3-4 hours (max. 6 hours) as required for an acute study. It is highly recommended to retain the wording of paragraph 7 to use bees from the hive and located on frames without brood (i.e. older than 2 days). Also, as honeybee foragers are expected to be the highest exposed caste it does not make sense to test the youngest bees which remain in the hive for the first 2-3 weeks. This is also relevant to consider for OECD TG 214 which simulates an exposure of a spray application.
 - Adult bees taken from the honey chamber are suitable.
9. **para 11 - OECD 213 and 214 are the only bee guidelines on laboratory level not asking for an analytical dose verification of the test substance solutions / feeding solutions. This requirement should be uniform across honeybee and bumblebee test guidelines. The wording should correspond to OECD Guideline 245 and OECD Guidelines 246 and 247.**
- Rational: In case the analytical recoveries do not meet the required range (typically within 20 % of nominal), an analytical verification of the test solutions / feeding solutions allows the calculation of nominal and analysed endpoints and verifies the correct preparation of the test solutions and dosing of the test organisms.**



Agreement justification:

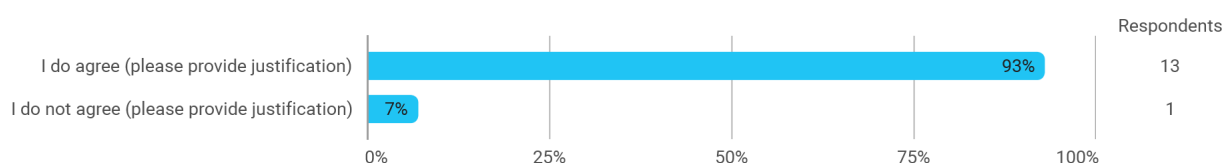
- Analytical verification of exposure is a critical component of toxicity testing
- should be uniform to the other bee studies, e.g. bumblebee
- Agreed, analytical verification of the lowest and highest concentration of feeding solutions (+ stock where applicable) would provide verification of actual exposure levels.
- That is right.
- Analytical verification should be part of the guideline

► Same a suggestion 1

Disagreement justification:

- For contact this is not relevant at all - given that the lab technician knows how to prepare test solutions. for oral testing, the feeding duration is only for 4h (if at all, mostly bees consume the provided food much faster), it is not necessary to prove analytical stability. This would more be a control of the ability of the lab technician.
- We disagree, as this would unnecessarily complicate the performance of such a test. As the exposure is very short and the entire test solution is consumed analytical verification should not be necessary for such a test system. In addition, in case clear precipitation of the test chemical is observed in the test solution by the qualified testing facility this is communicated and will result in lower test concentrations or a complete switch to another test item (i.e. formulated product instead of active ingredient).

10. para 11 - delete However, a concentration of the vehicle of up to 1% is generally appropriate and should not be exceeded.



Agreement justification:

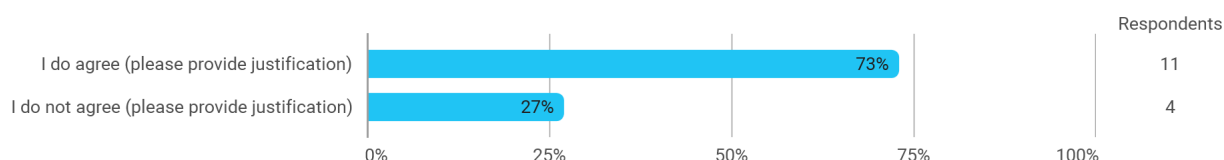
- concentration of the "vehicle" would depend on the vehicle and specifying a limit may be too detailed. Ideally, the solvent should be kept to a minimum but is largely dependent on control performance.
- e.g. for chronic honeybee tests concentrations of up to 5 % are valid
- higher concentrations of solvents should be permitted when the respective control group confirms the validity criterium. For example accord. to chronic bee studies up to 5 % v/v solvents (not acute bee toxic solvents) can be applied without any negative impact
- Agreed, this could be aligned with the wording in OECD GL 245: "The maximum acetone concentration in the final feeding solutions can be up to 5 %. Any other solvent, solubiliser or thickener can be used (e.g. to improve the homogeneity of the feeding solution during the 24 hours feeding interval) as long as the validity criterion for the control groups is met." Similar wording is used in OECD GL 246.
- Due to the solubility of the test item, 5% acetone or 2.5% DMSO are used for the oral test.
- Agree – however more information is needed on what to replace the sentence with – it should not just be deleted as we still need guidance on the maximum permissible solvent levels
- Agreed, However, each laboratory knows best which approach works for them. Therefore, we advise to allow for flexibility and not be too prescriptive.
- Agree – however more information is needed on what to replace the sentence with – it should not just be deleted as we still need guidance on the maximum permissible solvent levels

- ▶ If a solvent is needed a percentage of 1 % in the food is too low to test the intended doses. Acetone can be used up to 10 % in the food without having effects to the bees or the food consumption.

Disagreement justification:

- ▶ experience shows that the test vehicle is needed.

11. para 13 - Normally, five doses in a geometric series, with a factor not exceeding 2.5, and covering...



Agreement justification:

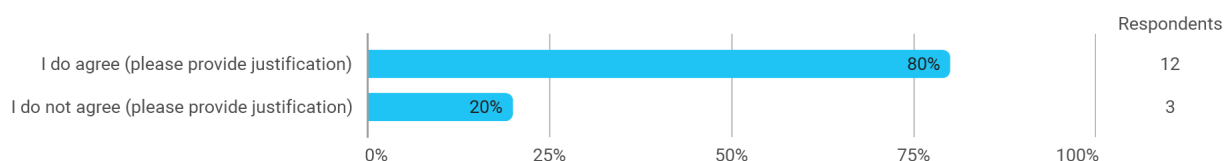
- ▶ correct
- ▶ a higher spacing factor can be a useful tool for some pesticides
- ▶ Agreed, a greater spacing factor allows for greater flexibility in allocating the dose range.
- ▶ Good idea
- ▶ Agreed - the spacing factor could be increased but usually factor of 2 is okay for an acute study
- ▶ Agreed, an increased spacing factor allows for greater flexibility. Should be adapted to the approaches used in the other bee test guideline/guidance documents.
- ▶ Agreed - the spacing factor could be increased but usually factor of 2 is okay for an acute study
- ▶ In some cases a wider factor is needed to meet the range of mortality

Disagreement justification:

- ▶ Suggestion 12 is more reasonable.
- ▶ From a stats perspective with the aim to calculate an LD_x a higher spacing factor doesn't make sense.
- ▶ Increasing the spacing factor might increase the imprecision of the measurement, instead more test doses should be used
- ▶ Spacing should be consistent for all TGs, terrestrial and aquatic. Harmonization in this area should be strongly considered.

12. para 13 - Verify if higher spacing factor (e.g., 2.5 - 3) can be used instead of 2.2 (also for other TG: 214, 246, 247).

Rational: Sometimes is more useful to use a higher factor



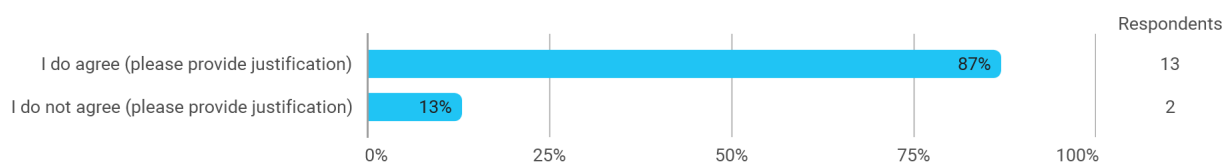
Agreement justification:

- ▶ Current factor of 2.2 seems inconsistent with other test guidelines and the proposed range provides greater flexibility.
- ▶ useful
- ▶ Agreed, a greater spacing factor allows for greater flexibility in allocating the dose range.
- ▶ Agreed but not essential
- ▶ Agreed, an increased spacing factor allows for greater flexibility. Should be adapted to the approaches used in the other bee test guideline/guidance documents.
- ▶ Agreed but not essential
- ▶ see above

Disagreement justification:

- ▶ From a stats perspective with the aim to calculate an LDx a higher spacing factor doesn't make sense.
- ▶ as above
- ▶ Same as Suggestion 11. Whatever is decided needs to be harmonized with all terrestrial and aquatic tests.

13. para 16 - Reduce to only one reference item group to be tested at a defined dose (e.g., 0.30 ug/bee) (also for OECD TG 214).



Agreement justification:

- ▶ Only one or two doses should be tested instead of generating an LD50. I would not add a specific dose but would more give a certain range within 50% mortality should be reached.
- ▶ sure, one dose sounds reasonable, but accord. to "Suggestion 3"
- ▶ Agreed, most laboratories have a good understanding of where to expect a significant effect from the toxic reference item and in an effort to reduce the number of bees to be sacrificed for testing, the toxic reference could be restricted to a single dose, as is the case with all other laboratory tests with bees. However, it is recommended to retain the option of the dose range so that laboratories can select the range that historically yielded a minimum 50% effect for them. In addition, by avoiding mentioning a specific dose, the toxic reference item

can be modified, as dimethoate may not be available for testing in the future and these dose rates are specific to dimethoate.

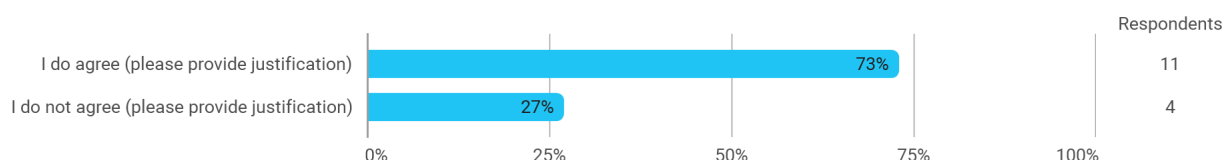
- ▶ Agreed – exact dose would need to be selected from historical data
- ▶ We agree. However, each laboratory knows best which test dose can elicit the effects needed to fulfil the necessary validity criteria. Therefore, we advise against stipulating a pre-defined dose as toxicity values generated in different laboratories can be subject to a certain level of variability.
- ▶ Agreed – exact dose would need to be selected from historical data
- ▶ One dose, but it should not be defined.
- ▶ Same response as with Suggestion 6.

Disagreement justification:

- ▶ Suggestion 15 is similar and seems more appropriate,
- ▶ only one reference item group reduces the sensitivity of bees

14. para 16 - The preferred toxic standard with known LD50 range is dimethoate. However, due to the non-approval of the active substance in the EU, the supply is subject to difficulties.

Rational: No change needed. However, alternative toxic standards and the overall need for reference substances should be discussed.



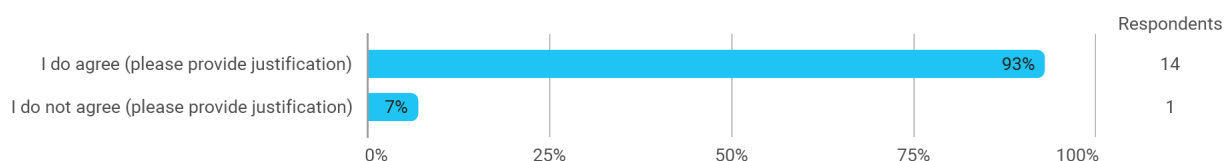
Agreement justification:

- ▶ Other possible toxic reference should be verified
- ▶ Agreed, alternative toxic reference items should be assessed. By keeping the wording on toxic reference item more flexible (i.e. stating that the validity criterion for the toxic standard is a minimum 50% effects instead of referring to a particular dose rate range).
- ▶ find alternative reference item
- ▶ Agree – need to find alternative toxic reference items and establish the expected LD50's/effect level
- ▶ We agree that alternative toxic reference items should be considered. Wording could be adapted that a suitable toxic reference substance (e.g. dimethoate) should be used at a dose that will fulfil the respective validity criteria.
- ▶ Agree – need to find alternative toxic reference items and establish the expected LD50's/effect level
- ▶ It is already difficult to obtain dimethoate in the EU

Disagreement justification:

- ▶ this seems more like an editorial comment rather than a recommended change. Concerns have been raised though regarding the availability of dimethoate. Whether those concerns should be included in the test guideline is uncertain.
- ▶ So far, it has never been a problem to get enough Dimethoate and currently no other toxic standard would be easier to get.
- ▶ This statement provides zero guidance. Either suggest a reference substance that is readily available or allow the lab to select one and stick with it for all of its reference tox doses.

15. para 16 - One dose of the reference substance leading to an expected mortality of $\geq 50\%$ at the end of the test period should be used to demonstrate the sensitivity of the bees and the reliability of the test system. Dimethoate can be used as reference substance. However, other toxic reference substances would be acceptable where sufficient data can be provided to demonstrate the expected sensitivity of the honey bees.



Agreement justification:

- ▶ Examples for further reference substances would be helpful.
- ▶ Suggested wording provides sufficient direction to document whether the study is able to detect the treatment effect without having an unreasonable focus on reference toxicant performance.
- ▶ Only one or two doses should be tested instead of generating an LD50. I would not add a specific dose but would more give a certain range within 50% mortality should be reached.
- ▶ sounds reasonable with respect to "Suggestion 14"
- ▶ Agreed, however, the wording should be slightly adapted to: One dose of the toxic reference substance leading to mortality of $\geq 50\%$ at the end of the test period should be used to demonstrate the sensitivity of the bees and the reliability of the test system. Dimethoate represents one possible toxic reference substance, however, other toxic reference substances would be acceptable where sufficient data can be provided to demonstrate the sensitivity of the exposed honey bees.
- ▶ We agree. However, each laboratory knows best which test dose can elicit the effects needed to fulfil the necessary validity criteria. Therefore, we advise against stipulating a pre-defined dose as toxicity values generated in different laboratories can be subject to a certain level of variability.

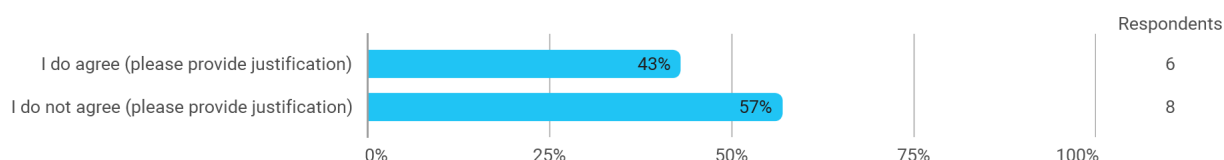
▶ Same as Suggestion 6.

Disagreement justification:

- ▶ should be uniform use of reference item

16. para 18 - The wording in paragraph 18 (test conditions) should be changed to: The bees should be kept in constant darkness (except during observation) under controlled climatic conditions at a target temperature of 33°C with maximum deviations of $\pm 2^\circ\text{C}$ and a relative humidity of 50 – 70 %. Short-term deviations (≤ 2 hours per day) from the recommended test conditions are unavoidable and should not affect the integrity or outcome of the test.

Rational: The required exact age of bees is not set in OECD 213 and 214 and it is simply stated, that the bees should be "young". To allow for a longer study duration and to simplify the fulfillment of the validity criteria, bees should be collected according to OECD 245. This however also leads to an adjustment of the test conditions, which should be identical to OECD 245 to account for the freshly hatched and acclimatized bees.



Agreement justification:

- Provides clarity and greater flexibility.
- would make it clearer.
- I am not an expert in this field. While the wording sounds reasonable, I will leave it up to the experts.

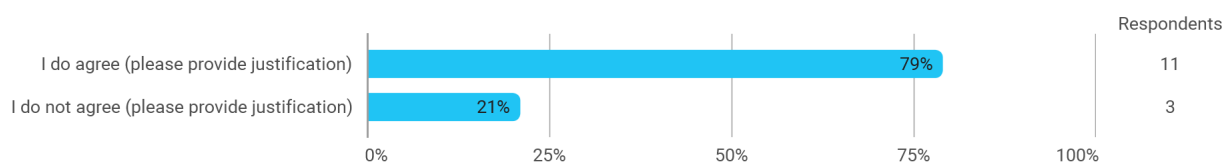
Disagreement justification:

- for testing forager bees, the climate conditions should be kept as they are actually stated in the guideline. Suggestion of the possibility of short-term deviations of at least 2 h per day should be added
- Disagreed. These test conditions are more appropriate for OECD GL 245, using young bees. Ambient room temperature (i.e. 25°C) is more representative of the conditions adult worker bees would be exposed to in the field, where an acute exposure would occur realistically. Suggest changing the wording to: The bees should be held in the dark (except during observation) in an experimental room at a temperature of $25 \pm 2^\circ\text{C}$. The relative humidity, normally around 50-70 %, should be recorded throughout the test. Short-term deviations (≤ 2 hours per day) from the recommended test conditions are unavoidable and should not affect the integrity or outcome of the test.
- Foraging bees should be kept $25 \pm 2^\circ\text{C}$, not like chronic oral test.
- Disagree – need more justification for changing the temperature of the study from 25 to 33°C – what is the driver for this? Is it to harmonize with the 10-day adult chronic study? The honeybee acute study has been run for a long time at 25 degrees so any increase in temperature would need to be justified. Consider a ring test comparing sensitivities of toxic references at different temperatures?
- We disagree. These test conditions reflect the requirements for OECD TG 245, HONEY BEE (APIS MELLIFERA L.), CHRONIC ORAL TOXICITY TEST (10-DAY FEEDING) which are

appropriate for young in-hive bees used in this test. Ambient temperatures (i.e. 25°C) are more suitable for adult forager bees which would be exposed to in the field, where an acute exposure would occur. We suggest changing the wording to: The bees should be held in the dark (except during observation) in an experimental room at a temperature of 25 ± 2 °C. The relative humidity, normally around 50-70 %, should be recorded throughout the test. Short-term deviations (≤ 2 hours per day) from the recommended test conditions are unavoidable and should not affect the integrity or outcome of the test.

- Disagree – These test conditions reflect the requirements for OECD TG 245, HONEY BEE (APIS MELLIFERA L.), CHRONIC ORAL TOXICITY TEST (10-DAY FEEDING) which are appropriate for young in-hive bees used in this test. Ambient temperatures (i.e. 25°C) are more suitable for adult forager bees which would be exposed to in the field, where an acute exposure would occur
- Why? A temperature of 25°C is suitable.

17. para 19 - After start of exposure to the test substance honey bees are observed for at least 48 hours. If test substance mortality increases by ≥ 10 % between 24 and 48 h, and furthermore between 48 and 72 h in one or more treatment group(s) the test should be extended up to 72 h or 96 hours, respectively.



Agreement justification:

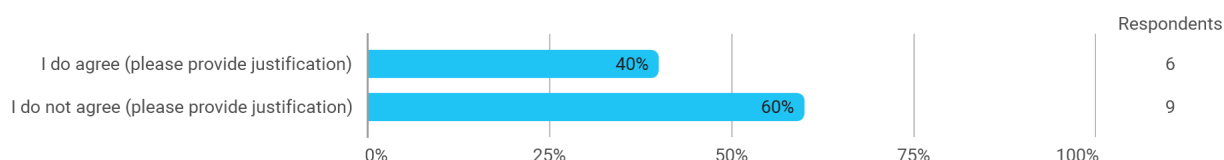
- Provides greater clarity.
- adaption of $\geq$ instead of $>$ would be useful.
- Agreed, however, the wording should be clarified slightly: After start of exposure to the test substance, honey bees are observed for at least 48 hours. If test substance mortality increases by ≥ 10 % between 24 and 48 h, and furthermore between 48 and 72 h in one or more test item treatment group(s) the test should be extended up to 72 h or 96 h, respectively.
- Yes, sentence of "in one or more treatment group(s)" is needed
- Agree – add clarification to wording
- We agree, However, we propose a slightly different wording: After start of exposure to the test substance, honey bees are observed for at least 48 hours. If test substance mortality increases by ≥ 10 % between 24 and 48 h, and furthermore between 48 and 72 h in one or more test item treatment group(s) the test should be extended up to 72 h or 96 h, respectively.
- Agree – add clarification to wording
- The prolongation can be done in two steps

Disagreement justification:

- ▶ either 48 hour or 96 hours when prolonged; because mortality can also rise up between the 72 h and 96 h assessments
- ▶ Same as Suggestion 2.

18. para 19 - It should be mentioned, that for microbials the test duration of several hours (max. 96 h) is too short, to assess pathogenic and infective properties of the test substance. Similar to OPPTS 885.4380 the test duration could be prolonged up to 30 days provided that the control mortality does not exceed 20%.

Rational: According to 238/2013 the toxic, infective and pathogenic properties of micro-organisms must be assessed.



Agreement justification:

- ▶ but strictly restricted to microbial pathogens. This is not necessary for plant protection products or other chemicals where acute mortality after exposure to a certain test item should be shown.
- ▶ for microbial should be added an information or linked reference

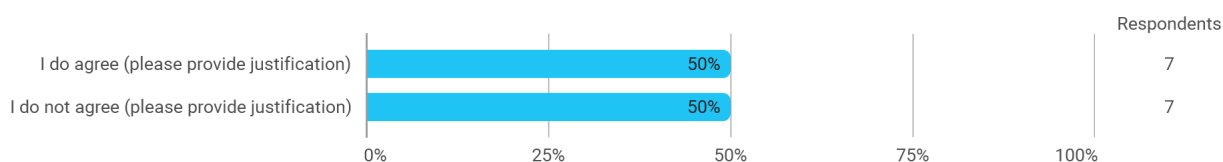
Disagreement justification:

- ▶ Microbial pesticides represent multiple challenges, and this guideline may not be an appropriate means of evaluating such test items. Otherwise, throughout the test guideline there would have to be caveats.
- ▶ Disagreed. This Guideline only addresses small molecule testing (pesticides and other chemicals) with adult worker honeybees. The US EPA microbial pesticide test guideline OPPTS 885.4380 observation period of 30 days is unrealistic, with control mortality exceeding 20% after around 20 days already. In addition, exposure verification, assessment of infectivity potential and limitations on test item concentration in the sugar diet causing palatability issues have to be assessed. A separate guidance on assessing microbial pesticides is needed, which cannot be covered by a guideline on acute exposure.
- ▶ 30 days is too long. Need a discussion about expose temperature .
- ▶ Disagree – we agree with the comment however there needs to be consideration as to whether this should be added to the test guideline. Changes to the methodology would need to be validated i.e. extending test duration to 30 days. There will likely be several instances where current test guidelines are not totally appropriate for testing microbials/biopesticides.
- ▶ We disagree. This Guideline only addresses small molecule testing (pesticides and other chemicals) with adult worker honeybees. The US EPA microbial pesticide test guideline OPPTS 885.4380 observation period of 30 days seems unrealistic. There would be the need to verify first that it is possible to keep bees in cages for 30 days and control mortality not

exceeding a potential validity criterium of 20%. In addition, exposure verification, assessment of infectivity potential and limitations on test item concentration in the sugar diet causing palatability issues have to be addressed. A separate guideline/guidance document on assessing microbial pesticides is needed, which cannot be covered by a guideline on acute exposure. This justification is also relevant for OECD TG 246 (BUMBLEBEE, ACUTE CONTACT TOXICITY TEST) and OECD TG 247 (BUMBLEBEE, ACUTE ORAL TOXICITY TEST).

- ▶ Disagree - this Guideline only addresses small molecule testing (pesticides and other chemicals) with adult worker honeybees. The US EPA microbial pesticide test guideline OPPTS 885.4380 observation period of 30 days seems unrealistic. There would be the need to verify first that it is possible to keep bees in cages for 30 days and control mortality not exceeding a potential validity criterium of 20%. In addition, exposure verification, assessment of infectivity potential and limitations on test item concentration in the sugar diet causing palatability issues have to be addressed. A separate guideline/guidance document on assessing microbial pesticides is needed, which cannot be covered by a guideline on acute exposure.
- ▶ For microbials a separate guideline is needed.
- ▶ is not it clear already that the whole series of TGs are for chemicals
- ▶ Thirty days is not considered an acute test. Suggestion should be to draft a new test guideline specifically for microbials to identify other unique aspects of the prolonged study.

19. para 19 - Existing text: The duration of the test is 48h. If mortality increases by more than 10 per cent between 24h and 48h, the test duration should be extended up to a maximum of 96h provided that control mortality does not exceed 10 per cent.
Suggested addition: The 'mortality increase' is calculated as the difference between the mortality level at 24 h and the level at 48 h. For example: if in a study mortality is 3.3% at 24 h and 6.7% at 48h, this is an increase of <10%, because $6.7 - 3.3 = 3.4\%$.
Rational: This section has caused some discussions in the past. Adding an explanation on how to calculate the percentage will make the guideline clearer.



Agreement justification:

- ▶ making an example would help to understand.
- ▶ Agree to add clarification to wording and how to calculate the mortality increase
- ▶ Okay, but furthermore add "per treatment group"

Disagreement justification:

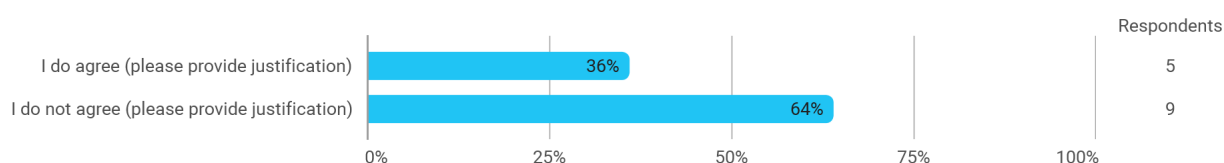
- ▶ Not sure whether this needs to be explained.
- ▶ do not understand this procedure, because the control validity is capped by 10 % for the whole test

- ▶ Disagreed, a guideline should be generic in its wording. Example calculations should not be necessary.
- ▶ I do not understand.
- ▶ We disagree. This guideline should be generic in its wording. Example calculations should not be necessary.
- ▶ Same as Suggestion 2.

20. para 19 - It is stated that the study duration should be extended to 96 h provided that the control mortality does not exceed 10 per cent. It should be clarified, that in case mortalities increase by more than 10 per cent after the assessment of 24 h and the study needs to be extended to 96 h, that it is mandatory to keep the mortalities below 10 per cent to provide a valid study even after 96 h.

Rational: Several studies with increasing mortalities and the subsequent need for prolongation of the study duration failed to provide a control mortality below 10 per cent at 96 h. Thus, it was justified by the laboratories, that effects for the assessment of 96 h could not be assessed, since mortalities did not meet the required range at 96 h, but the study is still valid for 48 h.

This is contradictory to the need of further assessments if mortalities tend to rise after the assessment of 24 h and results in an insufficient study outcome and false endpoints. Thus studies were then evaluated as invalid, as the required validity criteria could not be achieved.



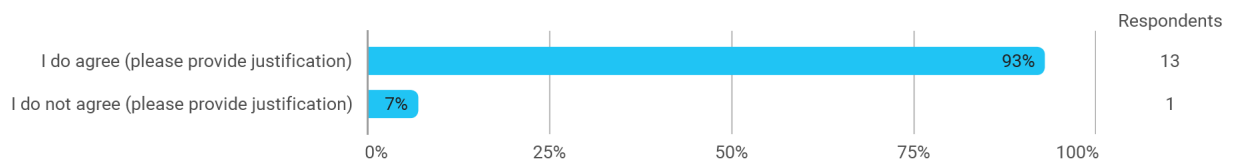
Agreement justification:

- ▶ would be helpful to make that point more clear.
- ▶ necessary to be clarified

Disagreement justification:

- ▶ Suggestion 17 provides sufficient clarity
- ▶ control validity of max. 10 % is absolutely feasible and should be fulfilled for validity of the study
- ▶ Disagreed. The current wording already implies this point when stating: “provided that control mortality does not exceed 10 per cent.”
- ▶ Disagree: We believe it is clear in the test guideline that the control validity criteria is 10% and the decision on whether to extend the study is based on test item data not control data.
- ▶ Prolongation should be done in two steps. First up to 72 hours and only in case of increased mortality between 48 and 72 h again up to 96 hours.
- ▶ Control mortality should be part of the acceptability criteria for the test. If the control mortality exceeds 10% at any time during the study, the test must be terminated.

21. para 20 - add text Mortality is recorded at 4 h after start of the test (start of feeding) and thereafter at 24 h and 48 h (i.e. deleted after given dose).



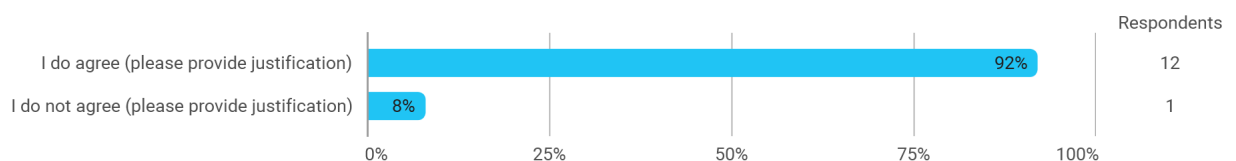
Agreement justification:

- ▶ Provides greater clarity
- ▶ would be helpful to make that point more clear.
- ▶ preisation of start of test to the start of feeding is needed
- ▶ Agreed. Suggest to slightly modify the wording to: Mortality is recoded at 4 h after start of the test (start of feeding) and thereafter at 24 h and 48 h after test start.
- ▶ We agree. We suggest adapting wording to: Mortality is recorded at 4 h after start of the test (start of feeding) and thereafter at 24 h and 48 h after start of feeding.
- ▶ "Start of feeding" is the benchmark for the timing of the assessments.
- ▶ This seems reasonable to discern the degree of sensitivity of the test organism to the test material

Disagreement justification:

- ▶ It depends on the result.

22. para 23 - If mortality occurs in the test substance treatment group which is statistically significantly higher compared to the relevant control group, a full dose-response study should be conducted.



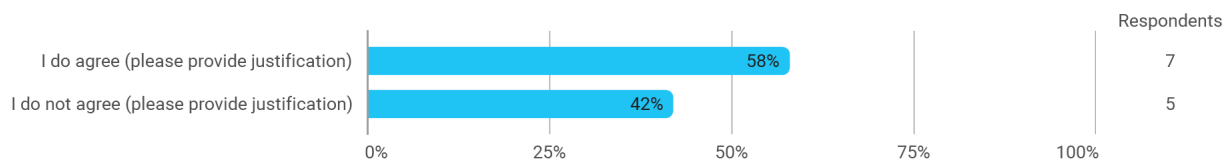
Agreement justification:

- ▶ Provides greater clarity
- ▶ Agreed. Suggest to slightly modify the wording to: If the mortality in the test substance treatment group is statistically significantly higher compared to the relevant control group, a full dose-response study should be conducted.
- ▶ Agree – clarification should be added to ‘if mortalities occur’ sentence. Agree with the approach of using statistics to compare control and treated to decide whether to do full dose response test
- ▶ The criterion for a dose-response test has to be specified.
- ▶ Dose response is critical if a limit rate cannot be determined.

Disagreement justification:

- If the test item has low toxicity but caused 20 % mortality. This is statistically significantly different to the control. I suppose no need to conduct a full dose response test.

23. para 22 - It would be useful to detail the requirements here.



Agreement justification:

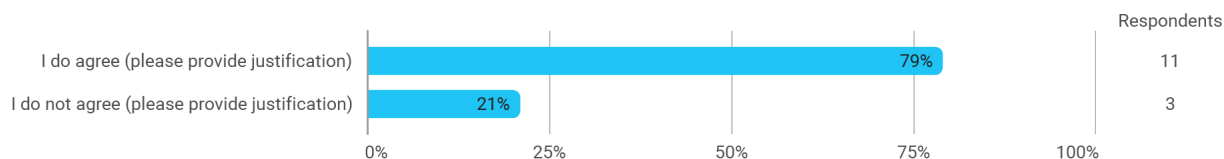
- like OECD 245
- see above

Disagreement justification:

- suggestion is vague
- not really needed because it's an artificial test and when differentiation that on a low level of approx. 2 different types of behavioural impairments
- Disagreed, a guideline should be generic in its wording. Exact descriptions of symptoms of behavioural abnormalities may differ between different labs.

24. para 23 - Not to limit the test dose to 100 ug/bee.

Rational: Low toxicity (i.e. 100 ug/bee) does not equal to low risk, especially for chemicals to which bees can be exposed at high level and repeatedly.



Agreement justification:

- While it is helpful to provide an upper limit, it is possible that estimated environmental levels may exceed the limit dose; therefore, it is helpful to have (to the extent possible) definitive LD50 values. Perhaps wording to the point that the limit dose selected should be reflective of environmentally relevant exposure levels.
- not really needed to limit a top dose to 100 µg ai/bee
- Agree – may want to test higher for non toxic a.i.'s?
- in some cases more than 100 µg a.i. has to be tested for the risk assessment
- very relevant point
- The highest rate could also be the limit of solubility of the test material

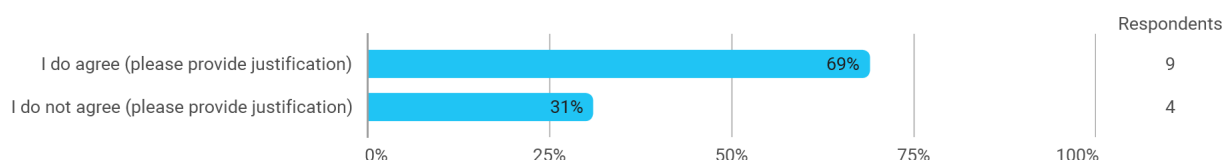
Disagreement justification:

- ▶ Disagree, the 100 µg a.i./bee is clearly a mere reference value here, whereby the LD50 is supposed to be greater than this value. This means that a higher dose could be used instead.
- ▶ Generally 100 µg a.i./bee or 200 µg product/bee are sufficient. However, tendence is higher than 100 µg.
- ▶ We disagree. The value 100 µg a.i./bee is considered as a reference value. This does not prohibit the use of a higher dose, if this is intended.

25. para 23 - The sentence "If mortalities occur, a full study should be conducted." should be specified:

If mortalities occur (≥ 10%), a full study should be conducted.

Rational: It is not specified and uncertain to what level occurring mortalities are acceptable in a limit test, before a full study is required. Considering the validity criterion for the control group of max. 10% average mortality, mortalities equal or greater than 10% compared to the control group should be regarded as significant to require a full (dose response) study.



Agreement justification:

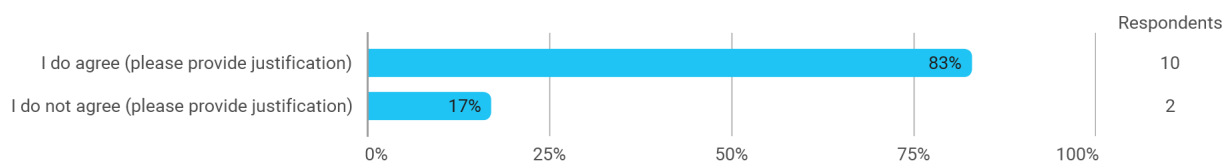
- ▶ correct, when a significant increase of bee mortality occurs
- ▶ Agreed. Suggest to slightly modify the wording to: If the mortality in the test substance treatment group is statistically significantly higher compared to the relevant control group, a full dose-response study should be conducted.
- ▶ should be specified
- ▶ See point above about using statistics – stats may be a better approach than an arbitrary value of 10% as it takes into account correcting for control mortality
- ▶ We agree. We suggest adapting wording to: If the mortality in the test substance treatment group is statistically significantly higher compared to the relevant control group, a full dose-response study should be conducted.

Disagreement justification:

- ▶ Is more appropriate the suggestion 23
- ▶ Covered by Suggestion 22.
- ▶ The statistical significance from the control should be the criterion
- ▶ The decision to jump from a limit test to a dose response test is statistically-based. This decision should be considered after review of the criteria set for aquatic tests. the approach should be the same.

26. para 24 - Consider the appropriate statistical tests.

Rational: Honey bees are housed in groups of 10 bees. Bees in one cage are not independent (a.o. because trophallaxis occurs), but they are currently treated as such in the statistical analysis. In reality, $n \neq 30$ but $n = 3$. We recommend that a statistical expert group reviews the test design and discusses how to address this pseudoreplication.



Agreement justification:

- ▶ totally agree
- ▶ Agreed, and this is already mentioned in the GL.
- ▶ missing NOED
- ▶ Agree – however this is a large exercise which would need to be harmonized across test guidelines and would require input from statisticians.
- ▶ Agreed, and this is already mentioned in the test guideline.
- ▶ Though a relatively meaningless statement without suggested methods.

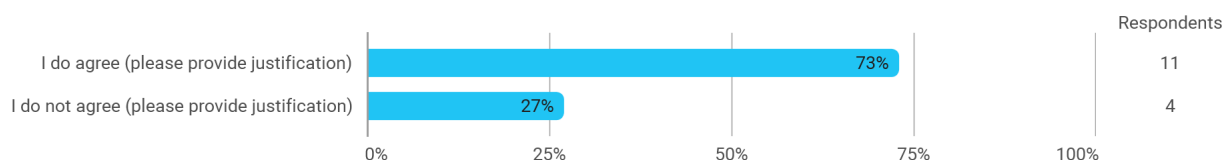
Disagreement justification:

- ▶ Vague
- ▶ Statistics is always data depended. There is not the one and only way forward when data are statistically analyzed. Therefore, it should be up to the study director to choose the right statistical test.

27. para 24 - Make it clear that Abbott's correction is always recommended or explain better when it is recommended or delete this recommendation. If kept, make it clear that which control should be considered (water control, solvent control, both, if both how).

Rational: Abbott's correction is to account for random mortality in the control and it assumes that the same random mortality have happened in the treatments, as well. The necessity of this approach might be questioned in a test when the control mortality is kept sufficiently low, like in this test. The test is valid only if the control mortality is below 10% and this is without differentiating whether 1, 2 or 3 bees died in the control. There are cases, when the mortality rate is lower in some treatments than in the control, through the assumption that the same random mortality have happened in the treatments clearly fails (and positive mortality is reported). Any corrections for effects in the control might be considered only in such test methods when this effect is normally considerable and would disturb the interpretation of the study. In this stable and repeatable laboratory test, the control mortality is always kept at low level. In addition, there is never a guarantee that certain effect that happened in a treatment is not due to the treatment itself. The involvement of solvent

control further complicates the case. Which control mortality should be considered as the reference and why?



Agreement justification:

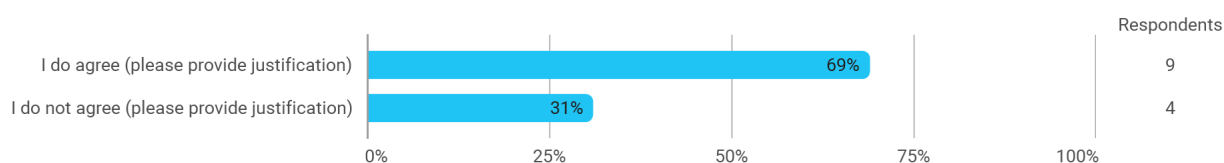
- ▶ Abbot's correction should be considered; however, the test guideline should not be prescriptive.
- ▶ Abbott correction is an unnecessary complication for such robust test than this one with control mortality < 10%
- ▶ like OECD 247
- ▶ Agreed - however this needs further discussion with statisticians and would apply across other test guidelines
- ▶ Abbott correction is an unnecessary complication for such robust test than this one with control mortality < 10%
- ▶ Agreed. Is it required, or not. Be clear.

Disagreement justification:

- ▶ is clear
- ▶ Disagreed – the guideline provides sufficient information on the fact that some level of correction for control mortality should be carried out.
- ▶ We disagree. Abbot's correction may not always be suitable. We advise to allow for flexibility and not be too prescriptive. The guideline provides sufficient information on the fact that some level of correction for control mortality should be carried out.
- ▶ Abbot correction should be mentioned as an option

28. paras 24 and 25 - LDx (e.g. LD10/LD20) might also be reported when possible to calculate.

Rational: LDx approach is more frequently used in regulatory context (though more for chronic tests); it might be considered in future.



Agreement justification:

- ▶ could be useful
- ▶ can be supplemented

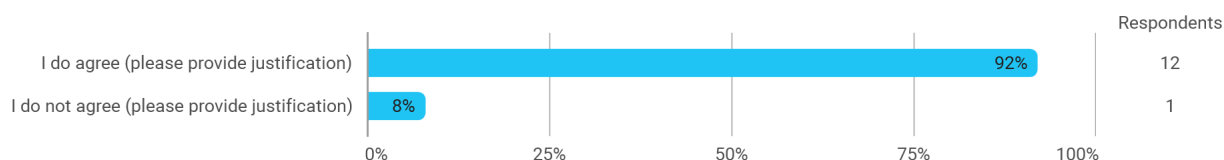
- ▶ Agreed, however, not all test designs (e.g. limit tests) may allow for the determination of LD10/20 values.
- ▶ when possible
- ▶ We agree. However, please consider that not all test designs (e.g. limit tests) may allow for the determination of LD10/20 values. Furthermore, we recommend adding this as an option and not a strict requirement, as this is not a requested endpoint in the EU. This justification is also relevant for OECD TG 246 (BUMBLEBEE, ACUTE CONTACT TOXICITY TEST) and OECD TG 247 (BUMBLEBEE, ACUTE ORAL TOXICITY TEST).

Disagreement justification:

- ▶ Not sure that is necessary to include this level of detail
- ▶ Disagree – not an EU data requirement and not necessary for an acute study – either use LD50 or NOED
- ▶ For acute toxicity tests, the LD50 is the applicable endpoint used in risk assessments. Therefore, additional calculations are not relevant for regulatory purposes.

29. para 25 - Report the slope in place of the graph with its 95% confidence interval (or both)

Rational: The use of the graphical presentation is limited, but the objective information on the slope is useful in regulatory context.



Agreement justification:

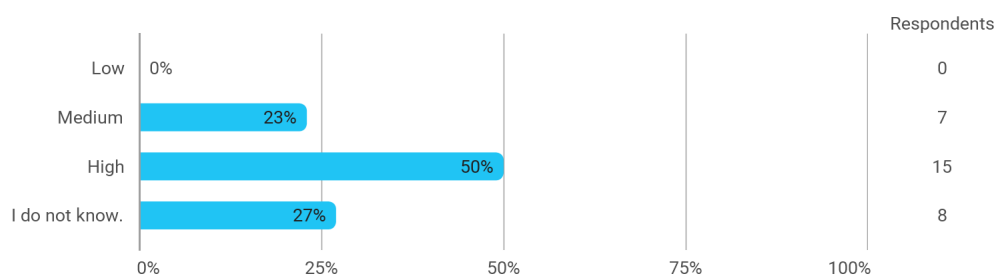
- ▶ It is helpful to understand the slope of the dose-response relationship
- ▶ both are already reported
- ▶ Agreed, the graph with dose-response curve along with its slope should be reported.
- ▶ Agree - report both
- ▶ We agree. No further justification necessary
- ▶ slope could be an important element for future RAs
- ▶ Useful information

Disagreement justification:

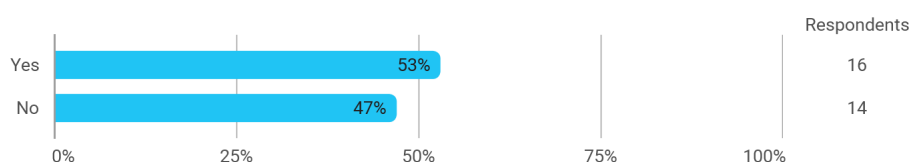
- ▶ In case of non linier regression the graph is not necessary

B.5 Test No. 214: Honeybees, Acute Contact Toxicity Test

Relevance of revision

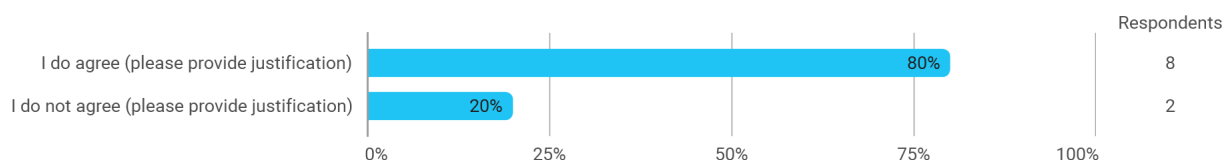


Interested to be part of an OECD Expert Group



The following suggestions have been collected:

1. **Add a section on analytical verification of the dose. The text could be the same as in OECD GL 246, contact toxicity to the bumble bee, paragraphs 19 and 20.**
Rational: The acute honeybee guidelines do not require analytical verification of the dose. All of the newer bee guidelines (OECD GL 237, 245, 246, 247 and GD 239 do include this requirement. Also in ecotoxicology tests with other organisms, analytical verification of the dose is common practice.



Agreement justification:

- ▶ Analytical verification of exposure is a critical component of the study.
- ▶ accordingly to other bee guidelines (bumblebee)
- ▶ Agreed, analytical verification of the lowest and highest concentration of feeding solutions (+ stock where applicable) would provide verification of actual exposure levels.
- ▶ yes, like OECD 246
- ▶ Agree – however need to consider all studies conducted to date that do not have analytical verification – we should ensure these studies are still valid.
- ▶ Makes sense

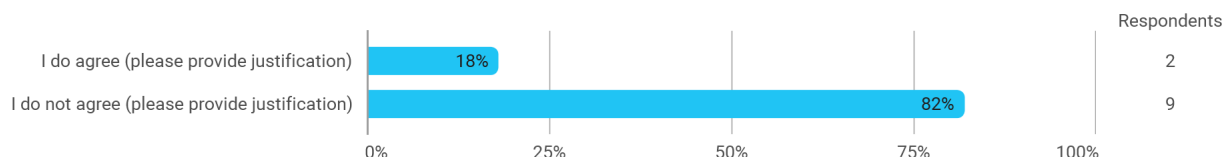
Disagreement justification:

- ▶ We disagree, as this would unnecessarily complicate the performance of such a test. As the exposure is very short and the entire test solution is consumed analytical verification should not be necessary for such a test system. In addition, in case clear precipitation of the test

chemical is observed by the qualified testing facility this is communicated and will have to result in lower test concentrations or a complete switch to another test item (i.e. formulated product instead of active ingredient).

- The cited paragraph is not suitable

- 2. It should be mentioned, that for microbials the test duration of several hours (max. 96 h) is too short, to assess pathogenic and infective properties of the test substance. Similar to OPPTS 885.4380 the test duration could be prolonged up to 30 days provided that the control mortality does not exceed 20%.**



Agreement justification:

- additional information regarding microbials testing is needed

Disagreement justification:

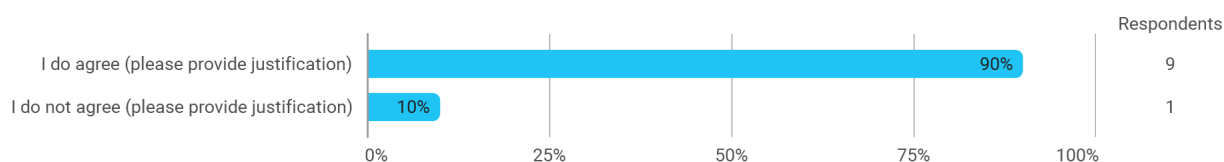
- Microbial pesticides represent multiple challenges and it is uncertain whether the various elements of this test guideline would need to flag each of these limitations.
- Disagreed. This Guideline only addresses small molecule testing (pesticides and other chemicals) with adult worker honeybees. The US EPA microbial pesticide test guideline OPPTS 885.4380 observation period of 30 days is unrealistic, with control mortality exceeding 20% after around 20 days already. In addition, exposure verification, assessment of infectivity potential and limitations on test item concentration in the sugar diet causing palatability issues have to be assessed. A separate guidance on assessing microbial pesticides is needed, which cannot be covered by a guideline on acute exposure.
- discussion is needed, exposure temperature, control mortality , ect.
- Disagree – we agree with the comment and it needs consideration but may not be appropriate to address within test guideline – this comment will likely apply to several test guidelines that were not originally designed with microbials in mind
- We disagree. This Guideline only addresses small molecule testing (pesticides and other chemicals) with adult worker honeybees. The US EPA microbial pesticide test guideline OPPTS 885.4380 observation period of 30 days seems unrealistic. There would be the need to verify first that it is possible to keep bees in cages for 30 days and control mortality not exceeding a potential validity criterium of 20%. In addition, exposure verification, assessment of infectivity potential and limitations on test item concentration in the sugar diet causing palatability issues have to be addressed. A separate guideline/guidance document on assessing microbial pesticides is needed, which cannot be covered by a guideline on acute exposure. This justification is also relevant for OECD TG 246 (BUMBLEBEE, ACUTE CONTACT TOXICITY TEST) and OECD TG 247 (BUMBLEBEE, ACUTE ORAL TOXICITY TEST).
- Disagree - this Guideline only addresses small molecule testing (pesticides and other chemicals) with adult worker honeybees. The US EPA microbial pesticide test guideline

OPPTS 885.4380 observation period of 30 days seems unrealistic. There would be the need to verify first that it is possible to keep bees in cages for 30 days and control mortality not exceeding a potential validity criterium of 20%. In addition, exposure verification, assessment of infectivity potential and limitations on test item concentration in the sugar diet causing palatability issues have to be addressed. A separate guideline/guidance document on assessing microbial pesticides is needed, which cannot be covered by a guideline on acute exposure.

- ▶ Microbials need a separate guideline
- ▶ the TG is for chemicals, exclusively
- ▶ Prepare a separate TG for biologicals

3. One or two test doses could be enough to demonstrate the sensitivity of the system in place of a full dose-response test for the toxic reference.

Rational: The aim of the toxic reference is to demonstrate the sensitivity of the system. To do that, a full series of doses to get a dose-response and derive a proper LD50 might not be necessary. Over the years, the method demonstrated a high level of repetability. OECD No. 237 (and OECD GD 239) require only one test dose.



Agreement justification:

- ▶ reasonable to save bees
- ▶ Agreed, most laboratories have a good understanding of where to expect a significant effect from the toxic reference item and in an effort to reduce the number of bees to be sacrificed for testing, the toxic reference could be restricted to a single dose, as is the case with all other laboratory tests with bees.
- ▶ We agree. However, each laboratory knows best which test dose can elicit the effects needed to fulfil the necessary validity criteria. Therefore, we advise against stipulating a pre-defined dose as toxicity values generated in different laboratories can be subject to a certain level of variability.
- ▶ One dose should be enough
- ▶ as for the oral
- ▶ If it can be shown over time that the same response is occurring from test to test.

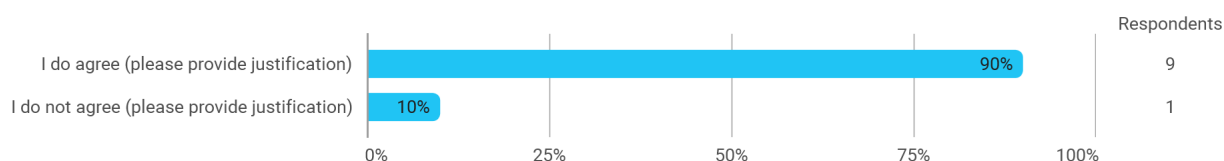
Disagreement justification:

- ▶ a full dose response test provides mor information

4. para 6 - It could be clarified whether the first validity criteria are for the negative control, the solvent control or both.

Rational: As the criteria currently reads, one could consider that averaging all controls is a good practice. That could lead validation of a test where a zero per cent

mortality was averaged with e.g. 17% mortality in the solvent control. However this test might not be considered as validity, since indicated adverse effects of the solvent used. It is advisable to clearly state in the TG that the role of using the solvent control is to demonstrate that the solvent did not influence the outcome of the test.



Agreement justification:

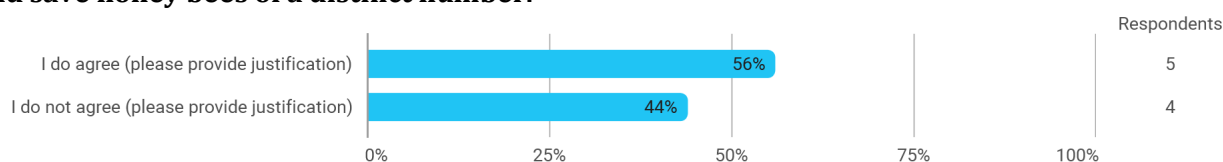
- ▶ Agreed, in the case where a solvent control is used the same validity criteria as for the control should be applied.
- ▶ need a clarification
- ▶ Agreed, no further justification needed.
- ▶ should be specified
- ▶ yes clarification is needed
- ▶ Validity criteria should be for both.

Disagreement justification:

- ▶ should be valid for all control groups

5. para 6 - Significant effects of the toxic standard should be observed at a concentration of 0.175 µg a.i./bee.

Rational: Due to numerous studies, the effects and outcome if the studies with the toxic standard are well known by the laboratories that perform acute bee studies. Given the fact that animal welfare is getting more and more important, this change would save honey bees of a distinct number.

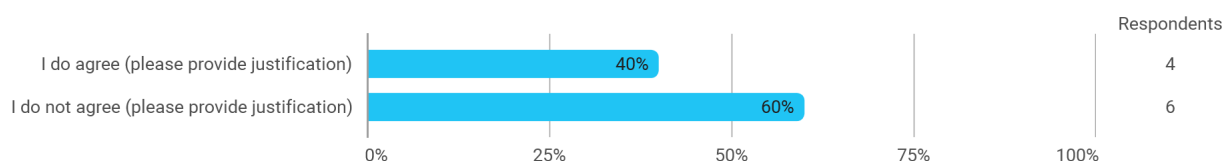


Agreement justification:

- ▶ great suggestion
- ▶ Agree that one rate of toxic standard could be used – the exact rate will need selecting using historical reference data
- ▶ Agree that one rate of toxic standard could be used – the exact rate will need selecting using historical reference data
- ▶ Same as Suggestion 3

Disagreement justification:

- ▶ Disagreed, the historical range for the 24hr oral LD50 for the toxic reference dimethoate is between 0.1 - 0.30 µg a.i./bee. Either this range should be specified or it should be clarified that a minimum of 50% effects is expected in the toxic reference group at the end of the test, independent of any particular dose.
 - ▶ We disagree with this comment. Toxicity values generated in different laboratories can be subject to certain levels of variability. In standard laboratory toxicity trials for honeybees (OECD 213 & 214, 1998) a range for the toxic standard dimethoate is given (LD50 (24h) of 0.1 - 0.35 µg a.i./bee). This range was proposed after 12 consecutive years of data collection (Gough et al., 1994). Therefore, the target of this Guideline should be to allow the test laboratory to identify an appropriate test dose which will fulfil the intended validity criterium and not to prescribe a dose that may not be suitable. Gough, H. J., McIndoe, E. C., & Lewis, G. B. (1994). The use of dimethoate as a reference compound in laboratory acute toxicity tests on honey bees (*Apis mellifera* L.) 1981–1992. *Journal of Apicultural Research*, 33, 119-125. doi:10.1080/00218839.1994.11100859
 - ▶ One dose: yes, but no recommended dose
6. **para 7 - The wording in paragraph 7 (collection of bees) should be changed to:**
Young bees (max. 2 days old) reared out from brood combs taken from queen-right colonies that have no symptoms of diseases and that have a known maintenance and physiological status history should be used for the test. No chemical substances (such as antibiotics, anti varroa treatments, etc.) should have been used in the hive for at least one month prior to the test. If one colony cannot provide the appropriate number of bees, comb(s) from several colonies may be used. In this case, it is ensured that the bees are equitably distributed across the treatments. Brood frames with capped cells that are expected to hatch on the same day can either be incubated in a climatic chamber or be kept without nurse bees in a worker excluder box within the hive until hatch. In the first case sufficient food supply should be ensured either by honey and pollen which is on the same brood comb or by an additional comb containing food. One day before the test starts, the bees can be collected from the combs and distributed into the test cages. Anesthetisation should be avoided during collection. Bees should be acclimated to test conditions for about one day (after a hatching period of one day). Bees are to be fed with sucrose solution ad libitum, but no additional feeding of pollen or water is necessary during the acclimation and test period. No starvation period is necessary before test start.



Agreement justification:

- Increases clarity
- Agree – text should be reviewed by laboratories conducting the work (if this comment did not originate from them)

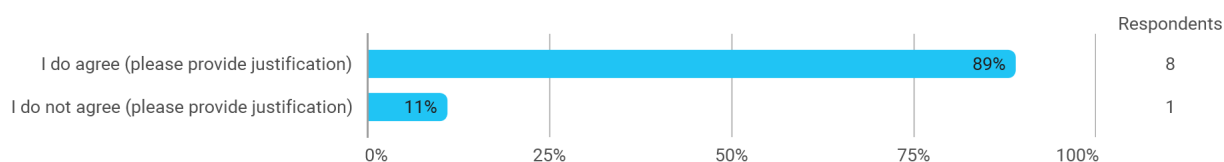
Disagreement justification:

- acute testing with forager bees, not nurse bees.
- Disagreed, this wording is taken from the OECD GL 245 assessing chronic exposure on young bees, where age synchronization and standardization are critical to maintain control validity criteria. However, experience has shown that young bees take longer to get used to the feeding system (i.e. syringes) and may not consume the treated diet within the 3-4 hours (max. 6 hours) that constitute an acute exposure scenario. It is highly recommended to retain the wording of paragraph 7 to use bees from the hive and located on frames without brood (i.e. older than 2 days).
- this is for chronic oral test
- We do not agree. This wording is taken from the OECD TG 245, HONEY BEE (APIS MELLIFERA L.), CHRONIC ORAL TOXICITY TEST (10-DAY FEEDING), where age synchronization and standardization are critical to achieve the validity criteria, especially for the control group. It is highly recommended to retain the wording of paragraph 7 to use bees from the hive and located on frames without brood (i.e. older than 2 days), as adult honey bee foragers are expected to be the highest exposed caste. Therefore, it does not make sense to test the youngest bees which remain in the hive for the first 2-3 weeks, as OECD TG 214 simulates an exposure of a spray application to bees outside of the hive.
- Adult bees taken from the honey chamber are suitable

7. **para 12 - The current wording specifies that 'appropriate control solutions should be prepared, i.e. where a solvent or dispersant is used to solubilise the test substance, two separate control groups should be used, one treated with water, and one treated with the solvent/disperser'.**

It is suggested to remove the reference to a water treated control, as without wetting agent the water will roll off the bee and therefore not provide an appropriate control comparison, a solvent/wetting agent control alone should be sufficient.

Rational: We suggest removal of a water treated control as without a wetting agent a droplet of water applied to the back of a bee will roll off and not provide a comparison to treatment



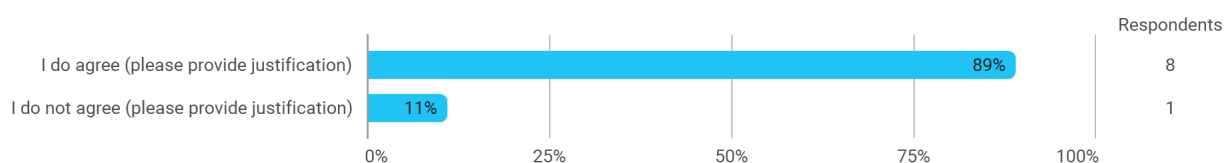
Agreement justification:

- ▶ totally agree to delete the water treated control in this system. the wetting agent control is sufficient as basis control. historical data also show that the wetting agent has no toxic effects at low concentration. Additionally, a uniform use of wetting agent (typ and concentration) should be specified in the guideline
- ▶ one treated water means water containing wetting agent
- ▶ Agree
- ▶ We agree. It would be good to clarify that the water control should always include a wetting agent to facilitate the spreading of the droplet on the bee thorax without it rolling off. A solvent control (which is different as the solvent facilitates the even spread of the test item in the solution) would also include the wetting agent.
- ▶ Agree
- ▶ A water weeding agent is mandatory

Disagreement justification:

- ▶ Disagreed. It would be good to clarify that the water control should always include a wetting agent to facilitate the spreading of the droplet on the hairy bee thorax without it rolling off. So that the specification water control group should therefore be used but including the wetting agent/dispersant, as to whereas the solvent control group refers to i.e., solvent, solubiliser including wetting agent/dispersant. Alternative proposal: 18. Appropriate control solutions should be prepared if a solvent, solubiliser, dispersant, etc. is used. In this case, two separate control groups should be used: one water control group including surfactant (= wetting agent), and one solvent control group (e.g., acetone) including surfactant (= wetting agent), at the same concentration as in the test chemical dose(s).

8. para 13 - Normally, five doses in a geometric series, with a factor not exceeding 2.5, and covering...



Agreement justification:

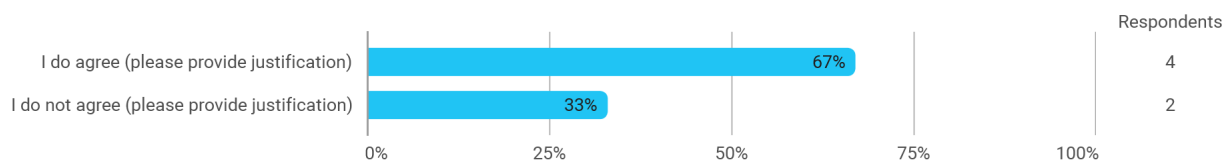
- ▶ agree, a higher spacing factor is useful for some tests
- ▶ Agreed, a greater spacing factor allows for greater flexibility in allocating the dose range.
- ▶ or not exceeding factor 3
- ▶ Agree consider larger spacing factor
- ▶ Agreed, an increased spacing factor allows for greater flexibility. Should be adapted to the approaches used in the other bee test guidelines/guidance documents.
- ▶ Agree consider larger spacing factor

- ▶ A factor of 2.5 is sometimes needed to cover the mortality range

Disagreement justification:

- ▶ Before spacing is agreed upon, strongly suggest comparing to spacing guidance provided in the aquatic TGs. They should all be harmonized.

9. para 16 - At least one dose within the reported LD50 (24h) range should be selected to show the sensitivity of the honeybees.



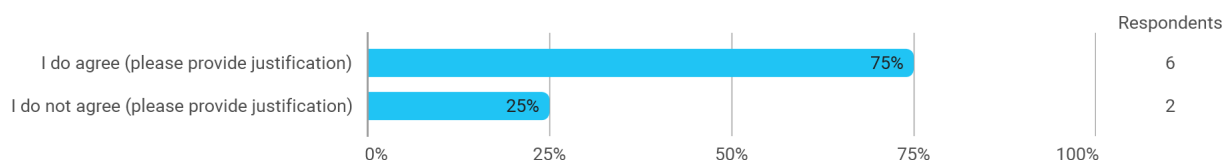
Agreement justification:

- ▶ Agreed, most laboratories have a good understanding of where to expect a significant effect from the toxic reference item and in an effort to reduce the number of bees to be sacrificed for testing, the toxic reference could be restricted to a single dose, as is the case with all other laboratory tests with bees.
- ▶ We agree. However, each laboratory knows best which test dose can elicit the effects needed to fulfil the necessary validity criteria. Therefore, we advise against stipulating a pre-defined dose as toxicity values generated in different laboratories can be subject to a certain level of variability.
- ▶ Same as Suggestion 5

Disagreement justification:

- ▶ one dose and a range sounds not really reasonable

10. para 16 - The preferred toxic standard with known LD50 range is dimethoate. However, due to the non-approval of the active substance in the EU, the supply is subject to difficulties.



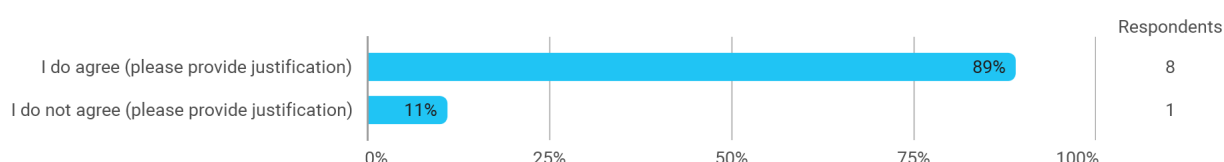
Agreement justification:

- ▶ Agreed, alternative toxic reference items should be assessed. By keeping the wording on toxic reference item more flexible (i.e. stating that the validity criterion for the toxic standard is a minimum 50% effects instead of referring to a particular dose rate range).
- ▶ need alternative reference item
- ▶ Agree – consideration needed to alternative reference items and expected LD50 ranges
- ▶ We agree that alternative toxic reference items should be considered. Wording could be adapted that a suitable toxic reference substance (e.g. dimethoate) should be used at a dose that will fulfil the respective validity criteria.

Disagreement justification:

- ▶ when one tox reference dose used at same wording as in the chronic or bumblebee testing guideline should be used ($\geq 50\%$)
- ▶ This statement provides no guidance. Either come up with another standardized reference toxicant or allow each testing facility to implement their own program with their own selected reference toxicant.

11. para 16 - One dose of the reference substance leading to an expected mortality of $\geq 50\%$ at the end of the test period should be used to demonstrate the sensitivity of the bees and the reliability of the test system. Dimethoate can be used as reference substance. However, other toxic reference substances would be acceptable where sufficient data can be provided to demonstrate the expected sensitivity of the honey bees.



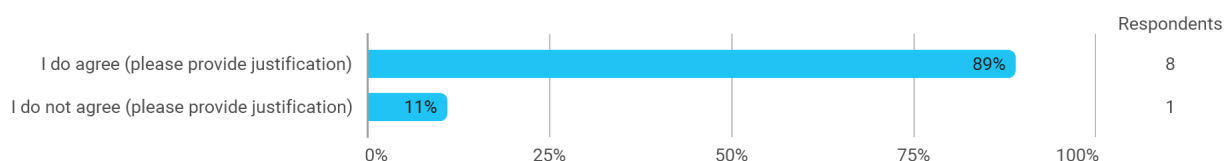
Agreement justification:

- ▶ great suggestion, but also specify the dose of the reference
- ▶ Agreed, however, the wording should be slightly adapted to: One dose of the toxic reference substance leading to mortality of $\geq 50\%$ at the end of the test period should be used to demonstrate the sensitivity of the bees and the reliability of the test system. Dimethoate represents one possible toxic reference substance, however, other toxic reference substances would be acceptable where sufficient data can be provided to demonstrate the sensitivity of the exposed honey bees.
- ▶ Agree -see above
- ▶ We agree. However, each laboratory knows best which test dose can elicit the effects needed to fulfil the necessary validity criteria. Therefore, we advise against stipulating a pre-defined dose as toxicity values generated in different laboratories can be subject to a certain level of variability.
- ▶ Agree -see above
- ▶ One dose of the reference item is enough to show the sensitivity
- ▶ Same as Suggestion 5. And agree with flexibility in selection of reference toxicant

Disagreement justification:

- ▶ not mortality of $\geq 50\%$. Certain range is needed.

12. para 17 - A volume of 2 µL (up to 5 µL) of solution containing the test substance at the suitable concentration should be applied with a suitable device (i.e. microapplicator, pipette..etc.) to the dorsal side of the thorax of each bee. Delete Other volumes may be used if justified.



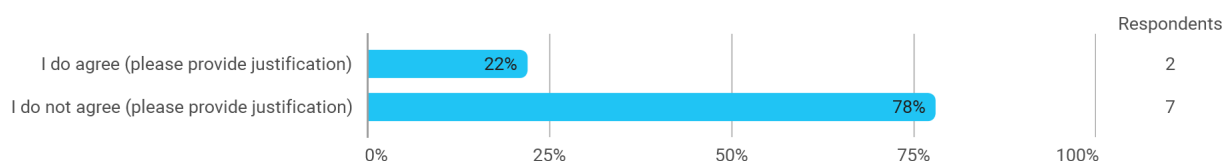
Agreement justification:

- ▶ great. higher volume of up to 5 µL/bee do not have a negative impact
- ▶ Agreed, slightly increased volumes will provide a better degree of volume dispersion on the bee thorax.
- ▶ A volume of 1 - 5 µL is preferable
- ▶ We agree. Slightly increased volumes will provide a better degree of volume dispersion on the bee thorax. This justification is also relevant for OECD TG 246 (BUMBLEBEE, ACUTE CONTACT TOXICITY TEST)
- ▶ 2 µL is the most common applied volume

Disagreement justification:

- ▶ Flexibility should be allowed if justified.

13. para 18 - The wording in paragraph 18 (test conditions) should be changed to: The bees should be kept in constant darkness (except during observation) under controlled climatic conditions at a target temperature of 33°C with maximum deviations of $\pm 2^\circ\text{C}$ and a relative humidity of 50 – 70 %. Short-term deviations (≤ 2 hours per day) from the recommended test conditions are unavoidable and should not affect the integrity or outcome of the test.



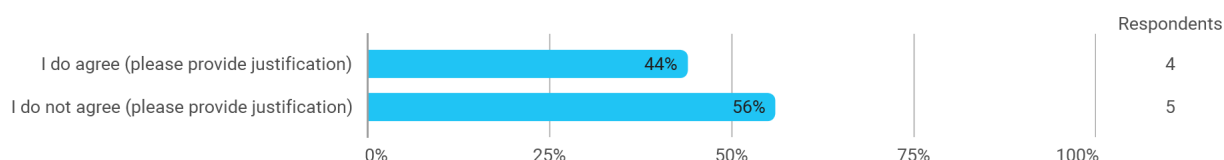
Disagreement justification:

- ▶ test conditions should be kept at 25 °C and 50-70%, but the possibility of short-term deviations should be added
- ▶ Target bees are foraging bees. 25°C is right.
- ▶ Disagree – need more justification for changing the temperature of the study from 25 to 33 – what is the driver for this? To harmonise with the 10 day adult chronic study? The study has been run for a long time at 25 so any increase would need to be justified. Consider a ring test comparing sensitivities at different temperatures?

- We disagree. These test conditions reflect the requirements for OECD TG 245, HONEY BEE (APIS MELLIFERA L.), CHRONIC ORAL TOXICITY TEST (10-DAY FEEDING) which are appropriate for young in-hive bees used in this test. Ambient temperatures (i.e. 25°C) are more suitable for adult forager bees which would be exposed to in the field, where an acute exposure would occur. We suggest changing the wording to: The bees should be held in the dark (except during observation) in an experimental room at a temperature of 25 ± 2 °C. The relative humidity, normally around 50-70 %, should be recorded throughout the test. Short-term deviations (≤ 2 hours per day) from the recommended test conditions are unavoidable and should not affect the integrity or outcome of the test.

- 25 °C is correct

14. para 19 - Existing text: The duration of the test is 48h. If mortality increases by more than 10 per cent between 24h and 48h, the test duration should be extended up to a maximum of 96h provided that control mortality does not exceed 10 per cent.
Suggested addition: The 'mortality increase' is calculated as the difference between the mortality level at 24 h and the level at 48 h. For example: if in a study mortality is 3.3% at 24 h and 6.7% at 48h, this is an increase of <10%, because $6.7 - 3.3 = 3.4\%$.
Rational: This section has caused some discussions in the past. Adding an explanation on how to calculate the percentage will make the guideline clearer.



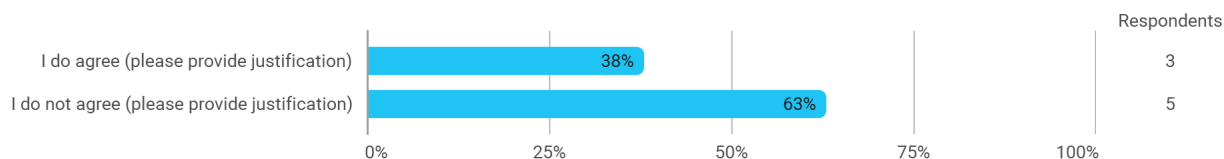
Agreement justification:

- Agree to add clarity on how to assess mortality increase
- Okay, but furthermore add "per treatment group"

Disagreement justification:

- the excising system is correct as it is
- Disagreed, a guideline should be generic in its wording. Example calculations should not be necessary.
- I am not sure what is mentioned.
- We disagree. This guideline should be generic in its wording. Example calculations should not be necessary.
- The study duration needs to be standardized, to avoid these discussions. There needs to be a level of consistency for all test guidelines, aquatic and terrestrial

15. para 19 - It is stated that the study duration should be extended to 96 h provided that the control mortality does not exceed 10 per cent. It should be clarified, that in case mortalities increase by more than 10 per cent after the assessment of 24 h and the study needs to be extended to 96 h, that it is mandatory to keep the mortalities below 10 per cent to provide a valid study even after 96 h.



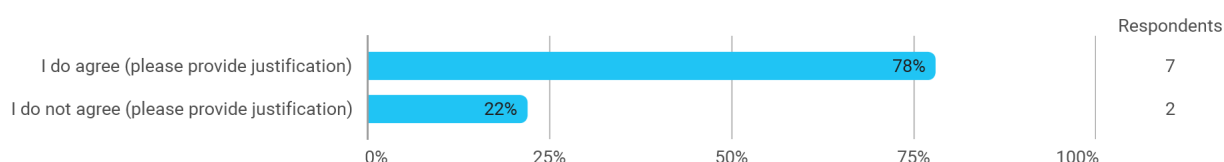
Agreement justification:

- Yes, control mortality should not exceed 10% at 96 hours.
- Not sure what this means

Disagreement justification:

- it's clear
- Disagreed. The current wording already implies this point when stating: “provided that control mortality does not exceed 10 per cent.”
- Prolongation should be done in two steps. First up to 72 hours and only in case of increased mortality between 48 and 72 h again up to 96 hours.
- Sane comment as Suggestion 15. And there needs to be an acceptability criterion for control mortality. If mortality exceeds 10%, the study is terminated.

16. para 19 - After start of exposure to the test substance honey bees are observed for at least 48 hours. If test substance mortality increases by $\geq 10\%$ between 24 and 48 h, and furthermore between 48 and 72 h in one or more treatment group(s) the test should be extended up to 72 h or 96 hours, respectively.



Agreement justification:

- Agreed, however, the wording should be clarified slightly: After start of exposure to the test substance, honey bees are observed for at least 48 hours. If test substance mortality increases by $\geq 10\%$ between 24 and 48 h, and furthermore between 48 and 72 h in one or more test item treatment group(s) the test should be extended up to 72 h or 96 h, respectively.
- "one or more treatment group(s)" is needed
- This is what is currently done, yes?
- This is what is currently done, yes?

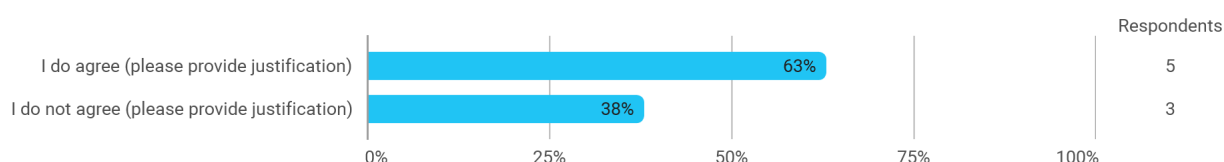
- add "at one or more test item dose levels"

Disagreement justification:

- neither 48 hour or 96 hours when prolonged; no need of 72 h tests. mortality also rises after 72 hours
- The study duration needs to be standardized, to avoid these discussions. There needs to be a level of consistency for all test guidelines, aquatic and terrestrial

17. para 21- Not to limit the test dose to 100 µg/bee.

Rational: Low toxicity (i.e. 100 µg/bee) does not equal low risk, especially for chemicals to which bees can be exposed at high level and repeatedly.



Agreement justification:

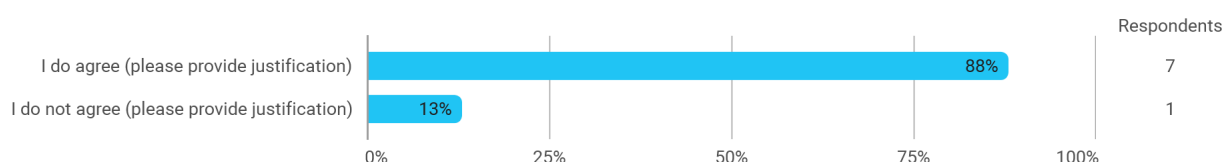
- Agree – it may be necessary to test higher for non toxic a.i.'s
- The limit of solubility should also be acceptable.

Disagreement justification:

- Disagree, the 100 µg a.i./bee is clearly a mere reference value here, whereby the LD50 is supposed to be greater than this value. This means that a higher dose could be used instead.
- It depends on the test item
- We disagree. The value 100 µg a.i./bee is considered as a reference value. This does not prohibit the use of a higher dose, if this is intended.

18. Report the slope in place of the graph with its 95% confidence interval (or both).

Rational: The use of the graphical presentation is limited, but the objective information on the slope is useful in regulatory context.



Agreement justification:

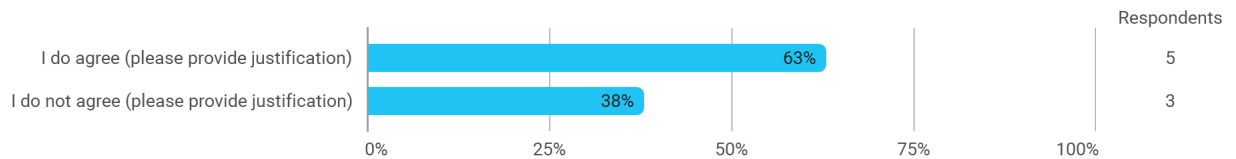
- both are already reported
- Agreed, the graph with dose-response curve along with its slope should be reported.
- Agree report both
- We agree to both. No further justification necessary.
- Agree report both

- Provides useful information.

Disagreement justification:

- It depends on a statistical method.

19. para 22 - It would be useful to detail the requirements here.



Agreement justification:

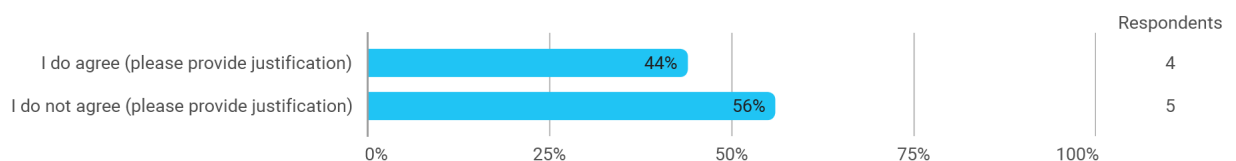
- like OECD 245

Disagreement justification:

- when specification then on a low level. due to this artificial test, it makes no difference to specify behavioural effects in a rich. Detail version
- Disagreed, a guideline should be generic in its wording. Exact descriptions of symptoms of behavioural abnormalities may differ between different labs.

20. para 22 - Make it clear that Abbott's correction is always recommended or explain better when it is recommended or delete this recommendation. If kept, make it clear that which control should be considered (water control, solvent control, both, if both how).

Rational: Abbott's correction is to account for random mortality in the control and it assumes that the same random mortality have happened in the treatments, as well. The necessity of this approach might be questioned in a test when the control mortality is kept sufficiently low, like in this test. The test is valid only if the control mortality is below 10% and this is without differentiating whether 1, 2 or 3 bees died in the control. There are cases, when the mortality rate is lower in some treatments than in the control, thrush the assumption that the same random mortality have happened in the treatments clearly fails (and positive mortality is reported). Any corrections for effects in the control might be considered only in such test methods when this effect is normally considerable and would disturb the interpretation of the study. In this stable and repeatable laboratory test, the control mortality is always kept at low level. In addition, there is never a guarantee that certain effect that happened in a treatment is not due to the treatment itself. The involvement of solvent control further complicates the case. Which control mortality should be considered as the reference and why?



Agreement justification:

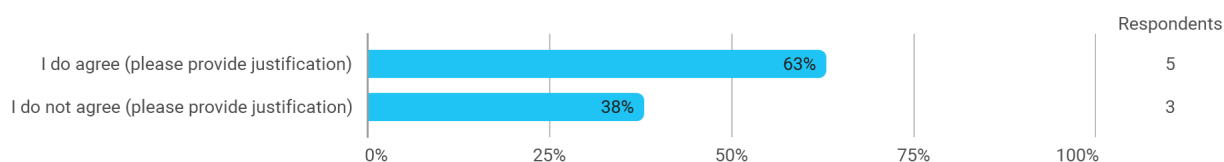
- ▶ like OECD 247
- ▶ Agree however this needs consideration from statistician

Disagreement justification:

- ▶ already clear
- ▶ Disagreed – the guideline provides sufficient information on the fact that some level of correction for control mortality should be carried out.
- ▶ We disagree. Abbot's correction may not always be suitable. We advise to allow for flexibility and not be too prescriptive. The guideline provides sufficient information on the fact that some level of correction for control mortality should be carried out.
- ▶ Abbot correction should be mentioned as an option
- ▶ If it is always recommended, then it is required. It is critical to be clear.

21. para 23 - LDx (e.g. LD10/LD20) might also be reported when possible to calculate.

Rational: LDx approach is more frequently used in regulatory context (though more for chronic tests); it might be considered in future.



Agreement justification:

- ▶ that can be included
- ▶ Agreed, however, not all test designs (e.g. limit tests) may allow for the determination of LD10/20 values.
- ▶ yes, if possible
- ▶ We agree. However, please consider that not all test designs (e.g. limit tests) may allow for the determination of LD10/20 values. Furthermore, we recommend adding this as an option and not a strict requirement, as this is not a requested endpoint in the EU. This justification is also relevant for OECD TG 246 (BUMBLEBEE, ACUTE CONTACT TOXICITY TEST) and OECD TG 247 (BUMBLEBEE, ACUTE ORAL TOXICITY TEST).

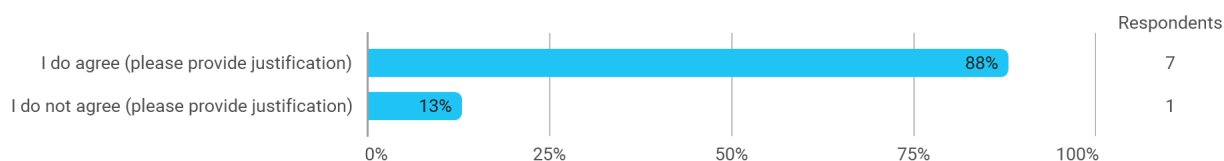
Disagreement justification:

- ▶ Disagree – not a data requirement, not needed for acute study, LD50 and NOED should suffice
- ▶ The LD50 is the relevant endpoint that is used in risk assessment for acute testing. Therefore, any other metric is not relevant, from a regulatory perspective.

22. para 24 - Consider the appropriate statistical tests.

Rational: Honey bees are housed in groups of 10 bees. Bees in one cage are not independent (a.o. because trophallaxis occurs), but they are currently treated as such

in the statistical analysis. In reality, $n \neq 30$ but $n = 3$. We recommend that a statistical expert group reviews the test design and discusses how to address this pseudoreplication.



Agreement justification:

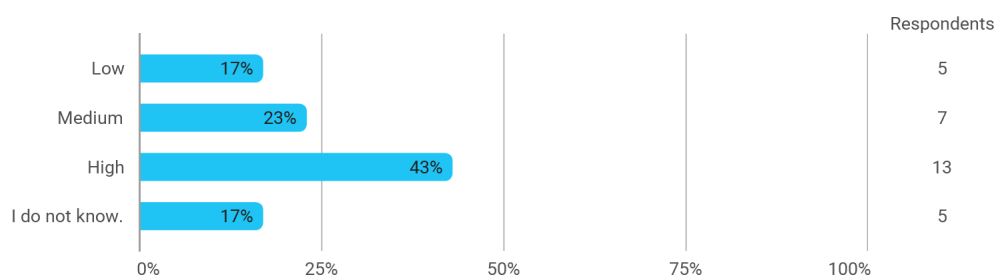
- ▶ information on appropriate stat tests are needed
- ▶ Agreed, and this is already mentioned in the GL.
- ▶ NOED, etc.
- ▶ See other comments on statistics – would need effort across all test guidelines
- ▶ We agree. This is already mentioned in the test guideline.
- ▶ See other comments on statistics – would need effort across all test guidelines

Disagreement justification:

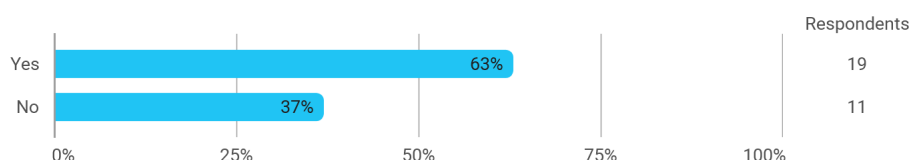
- ▶ This is a meaningless statement without recommendations of relevant statistical methods.

B.6 Test No. 237: Honey Bee (*Apis Mellifera*) Larval Toxicity Test, Single Exposure

Relevance of revision

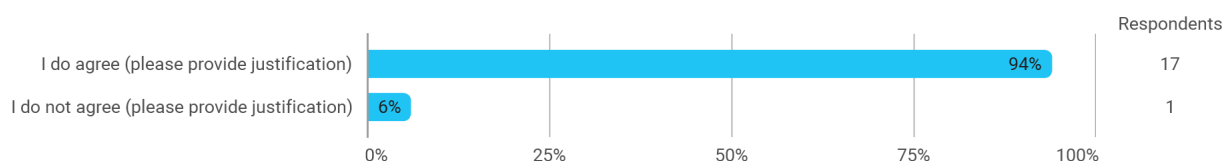


Interested to be part of an OECD Expert Group



The following suggestions have been collected:

1. It should be made clear what ‘across replicates’ means exactly. Please note that there are many comments for TG 213/214 that would be relevant for this TG as well. Rational: It is unclear if ‘across replicates’ mean the average of all replicates or it refers to each individual replicates.



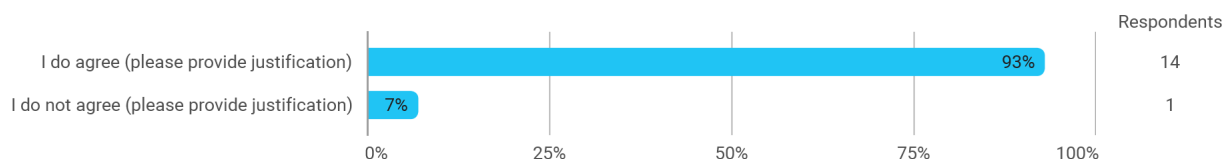
Agreement justification:

- ▶ very important point
- ▶ There are comments on OECD 213 and 214 that would be relevant to TG 237
- ▶ I agree to the suggestion. Not each replicate should be analysed individually, because with 12 larvae/rep. 15% mortality are 1.8, which means that only 1 larve per repl. is allowed to die.
- ▶ I is an important one, but had already sorted out (at least for GD239)
- ▶ Agreed, yet clarification is needed on what comments from 213/214 are relevant for this TG that may not currently be listed here.
- ▶ The way of interpretation of "across replicates" ("average" or "per replicate") has an immense influence on the validity of the test, therefore clarification is needed
- ▶ across replicates is needed
- ▶ important to avoid misinterpretation
- ▶ We agree. No additional justification needed.
- ▶ see next comment
- ▶ It is an important one, but had already sorted out (at least for GD239)
- ▶ This point needs to be clarified.

2. **para 7 - Change the sentence "in the control plate(s), cumulative larval mortality from D4 to D7 should be $\leq 15\%$ across replicate" to the sentence "in the control(s), the average cumulative larval mortality from D4 to D7 should be $\leq 15\%$ " this is valid also for the GD239.**

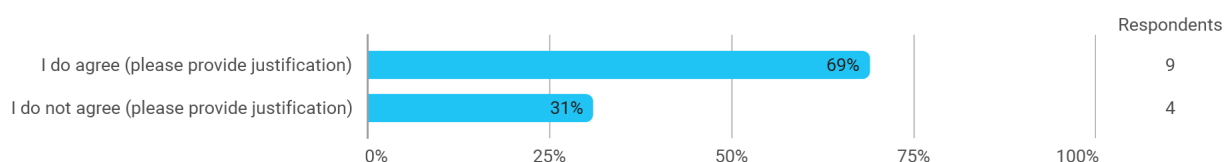
Rational: The wording suggests that the criteria are to be met 'across' replicates typically by using the average across control replicates while missing the word 'average' over replicates allows the interpretation that the criteria are to be met for each individual replicate.

For each ecotox test, the comparison between control and test item treatments is always performed by using an average across all replicates: there is no knowledge of OECD protocols testing the toxicity of chemicals to non-target organisms that would apply control validity criteria to each individual replicate of the control group.



Agreement justification:

- ▶ Agreed, yet wording should match what is being proposed for OECD GD 239 and be written as, “in the control(s), the average cumulative larval mortality as the mean across all replicates from D4 to D7 should be $\leq 15\%$ ”.
 - ▶ It can be shown by calculations using binomial formulas, that relating the 15% criterion to single replicates would result in a considerable proportion of the tests being invalid, depending on the assumed natural mortality rate of the larvae: - if natural (random) mortality rate 5%, then statistically it can be expected that 31% of the tests are invalid (compared to 0.8% if validity criterion is related to sum of all replicates - if natural (random) mortality rate is 10%, then statistically 71% of the tests are expected to be invalid (compared to 14% if validity criterion is based on the sum of all replicates).
 - ▶ important to avoid misinterpretation
 - ▶ Agree. Please be aware that there is an update available for OECD GD 239 (July 2021) where it states the following on the title page: “Following several questions raised after publication, the OECD would like to clarify that the validity criteria mentioned in paragraph 7 «across all replicates» should be understood as the average value across all replicates, a replicate being defined as the number of larvae originating from the same colony.”
 - ▶ should be adapted to GD 239
 - ▶ More concise and consistent with TG239.
3. **para 21 - Change the sentence: "however, deviations are allowed, but temperature should not be lower than 23°C or higher than 40°C, and these deviations should not last more than 15 minutes once every 24 hours." to "temporary deviations are UNAVOIDABLE (e.g., during treatment, feeding and assessments). These deviations need to be evaluated (e.g., 2 hours out of the range could not affect the outcome of the test). However, temperature should not be lower than 23°C or higher than 40°C for more than 30 minutes once every 24 hour" also for OECD GD239.**
- Rational: Temporary deviation always occur during treatment, feeding, assessment when the desiccator is opened and closed because the system needs time to return to the required climatic conditions since it is a hermetically sealed system. The humidity within the desiccator depends on a saturated salt solution and the temperature depends on heat exchange through the walls. Also for OECD GD239**



Agreement justification:

- ▶ very important point
- ▶ An evaluation of the deviations caused by feeding or assessment should not be "evaluated" in each individual study. The duration of the deviation should be increased to 30 minutes, because with a climate interval of 1 measurement per 10 minutes a second outlier in a row would be too much.

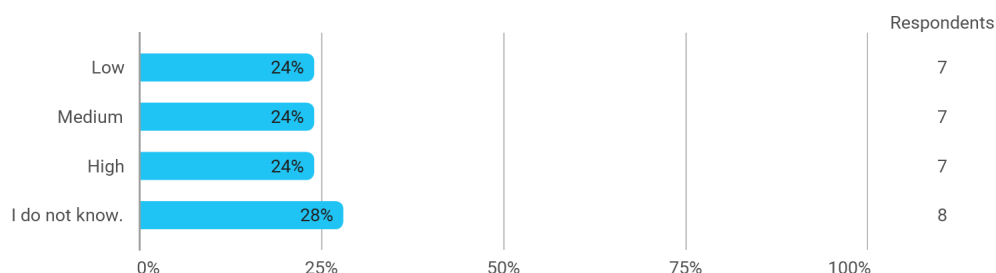
- ▶ However, it must be added that it must be evaluated only after a certain period of time. Simultaneous to the initial sentence.
- ▶ it is clear to understand
- ▶ Agree in principle – suggest this wording is reviewed by those regularly conducting the studies at laboratories
- ▶ Agree in principle – suggest this wording is reviewed by those regularly conducting the studies at laboratories
- ▶ climatic conditions should be specified
- ▶ Provides better guidance to assess and report potential impacts on the study.

Disagreement justification:

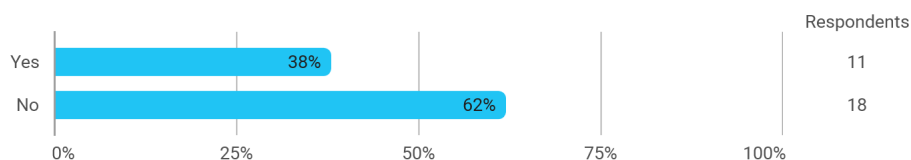
- ▶ These deviations need to be evaluated ... if they last more than 30 minutes
- ▶ We disagree. The OECD 237 should include “However, temperature deviations are allowed, but should not be lower than 23°C or higher than 40°C for more than 30 minutes once every 24 hours. All temperature deviations must be recorded”. The other recommended text here (along with the examples) is unnecessary. This text should match what is written in OECD GD 239.
- ▶ Instead of 'unavoidable', the text should use 'tolerated'

B.7 Test No. 245: Honey Bee (*Apis Mellifera* L.), Chronic Oral Toxicity Test (10-Day Feeding)

Relevance of revision



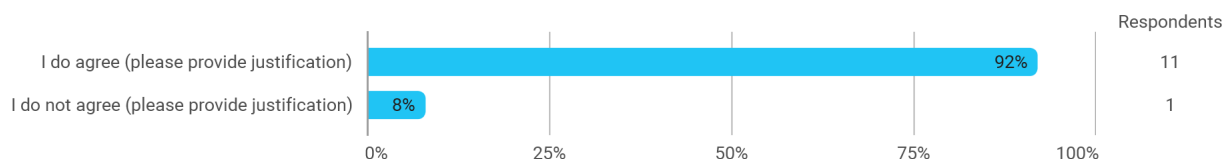
Interested to be part of an OECD Expert Group



The following suggestions have been collected:

1. It should be made clear what ‘across replicates’ means exactly.

Rational: It is unclear if ‘across replicates’ mean the average of all replicates or it refers to each individual replicates.



Agreement justification:

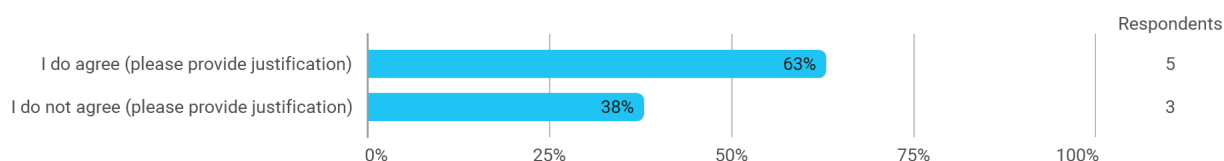
- ▶ but already sorted out
- ▶ Agreed: Clarification on this point is welcome, however, as for OECD GD 239 this could be added to the cover page and does not require a formal update or revision of the test guideline.
- ▶ The way of interpretation of "across replicates" ("average" or "per replicate") has an immense influence on the validity of the test, therefore clarification is needed. It can be shown by calculations using binomial formulas, that relating the 15% criterion to single replicates would result in a considerable proportion of the tests being invalid, depending on the assumed natural mortality rate of the larvae: - if natural (random) mortality rate 5%, then statistically it can be expected that 31% of the tests are invalid (compared to 0.8% if validity criterion is related to sum of all replicates - if natural (random) mortality rate is 10%, then statistically 71% of the tests are expected to be invalid (compared to 14% if validity criterion is based on the sum of all replicates).
- ▶ important to avoid misinterpretation; must read "mean of all replicates"
- ▶ We agree. Please be aware that there is an update available for OECD GD 239 (July 2021) where it states the following on the title page: "Following several questions raised after publication, the OECD would like to clarify that the validity criteria mentioned in paragraph 7 «across all replicates» should be understood as the average value across all replicates, a replicate being defined as the number of larvae originating from the same colony."
- ▶ but already sorted out
- ▶ This needs to be clarified.

Disagreement justification:

- ▶ "across replicates" is not necessary for the chronic oral test

2. para 31 - Consider the appropriate statistical tests.

Rational: Honey bees are housed in groups of 10 bees. Bees in one cage are not independent (a.o. because trophallaxis occurs), but they are currently treated as such in the statistical analysis. In reality, $n \neq 30$ but $n=3$. We recommend that a statistical expert group reviews the test design and discusses how to address this pseudoreplication.



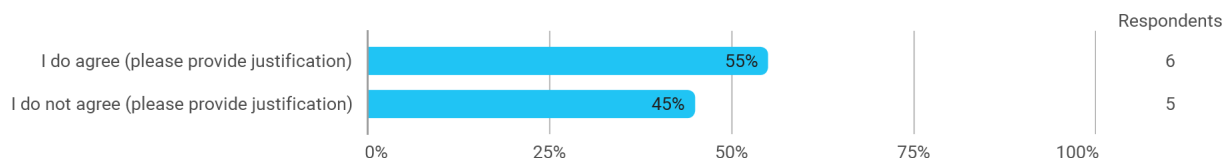
Agreement justification:

- ▶ How is the test item compared to the control (water, solvent or pooled)?
- ▶ Agree in principle however this would require an extensive harmonized approach across test guidelines with advice from statisticians

Disagreement justification:

- ▶ We disagree. Para 31 cites examples of appropriate statistical tests and how to correct for control mortality if necessary. Further guidance on the use of statistical tests in ecotoxicology is available in OECD Series on Testing and Assessment, No. 54. Consequently, a formal update or revision of the test guideline is not applicable.
- ▶ This is a meaningless statement with recommending statistical methods.

3. General suggestion to express toxicity relative to bodyweight (mg/kg) to account for the differences in sizes between different bee species. For meaningful comparisons to be made between species it is suggested to move away from expressing toxicity on a per bee basis and move towards a size/weight corrected expression, we therefore suggest that body weight measurements are taken as is currently recommended in bumblebee test guidelines OECD 246 and OECD 247



Agreement justification:

- ▶ BW data helpful (therefore not disagreed), however this issue is a general one and belongs to risk assessment, not for a specific TG for HB
- ▶ Agree, but depends on what endpoint is suggested for risk assessment
- ▶ BW data helpful (therefore not disagreed), however this issue is a general one and belongs to risk assessment, not for a specific TG for HB
- ▶ I am not an expert in this area. As someone who prepares risk assessments, it seems reasonable to make this switch, relative to body weights. This would be consistent with other areas of the regulatory risk assessment world. But, I will leave this decision to the experts.

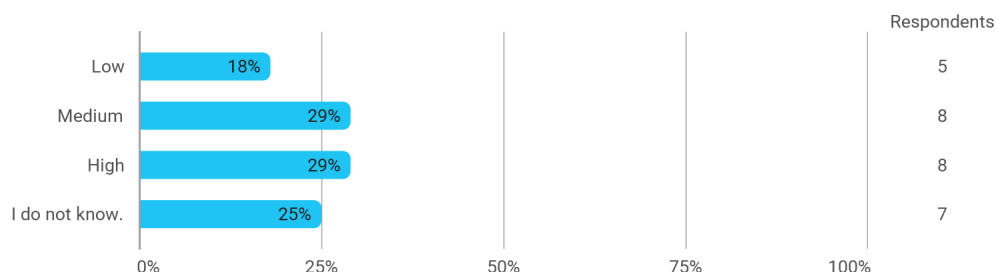
Disagreement justification:

- ▶ the weight for bumblebees is used only to eliminate the very small and very big bb, and the endpoints are expressed as mg/BB

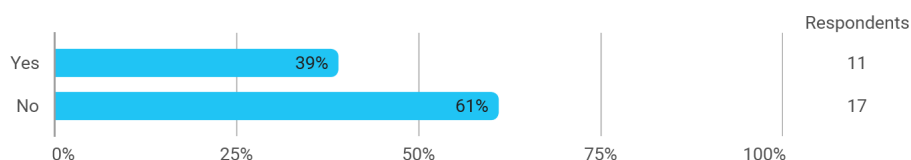
- Disagree: The bumblebee test guidelines OECD 246 and OECD 247 specify the need to make body weight measurements because the size of bumble bee individuals can vary greatly. The weighed bumble bees are then randomly assigned to test groups to ensure that there is no size bias across treatment doses. This is not necessary for honey bees as are they are all of very similar size and can be used without the need to pre-weigh each bee. Honey bees are group housed in the test units (unlike bumble bees which are held individually) and it would not be possible to know how much each individual bee consumes and caging honey bees individually is undesirable due to their social behaviour. Furthermore, weighing the honey bees prior to test would also increase the handling and stress of young bees leading to possibly higher control mortalities.
- In case of the bumble bees the body weight varied very wide. However, honey bees are not so much different of the body weight. No need to express toxicity to weight body.
- We disagree. The bumblebee test guidelines OECD 246 and OECD 247 specify the need to make body weight measurements because the size of bumble bee individuals can vary greatly. The weighed bumble bees are then randomly assigned to test groups to ensure that there is no size bias across treatment doses. This is not necessary for honey bees as are they are all of very similar size and can be used without the need to pre-weigh each bee.

B.8 Test No. 246: Bumblebee, Acute Contact Toxicity Test

Relevance of revision

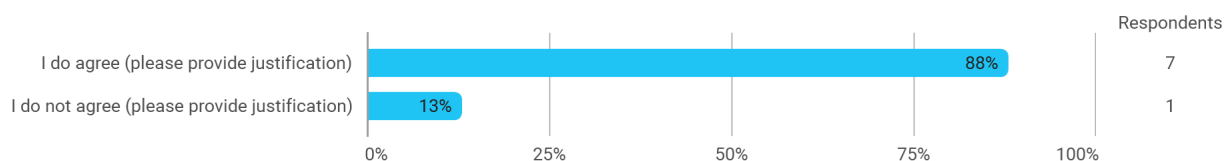


Interested to be part of an OECD Expert Group



The following suggestions have been collected:

1. Please see the comments submitted for TG 214 - They are also relevant for this TG. However, it was not possible in the survey to enter comments for several TG.



Agreement justification:

- Agreed that there are suggestions for TG213 and TG214 which are relevant.

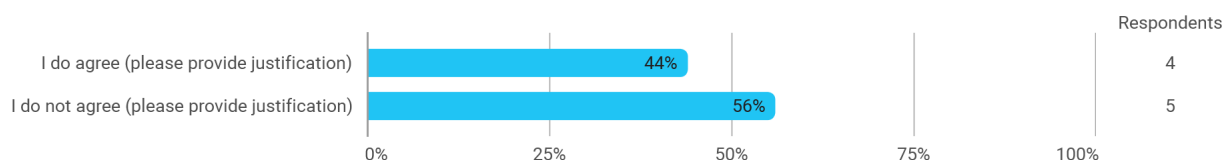
- Agreed. This action takes a lot of time and has no relevance for the study outcome or does influence the results in any way.
- Not sure what answer you are expecting for with this comment.

Disagreement justification:

- We disagree, not necessary as this guideline was just published very recently

2. para 14 - Delete the sentence : “Bumblebees used in the test are weighed individually”, also for OECD TG247.

Rational: no range of weight is provided but only a qualitative indication (very small and particularly very large bumblebees) so this operation have no sense, probaly it is a refuse of the first version of the draft guideline. Also for OECD TG247



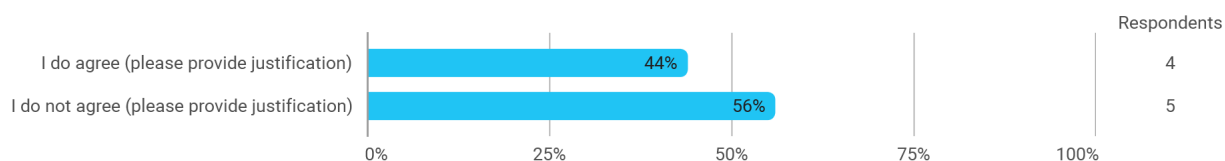
Agreement justification:

- Agreed. This action takes a lot of time and has no relevance for the study outcome or does influence the results in any way.
- Appropriate to take either a measurement from entire group to minimize over-handling or from a sub-population of the bees selected but not used in the actual test.

Disagreement justification:

- Unlike honey bees, bumble bees can vary widely in weight; having such data provides a means of assessing the extent to which weights actually varies
- Body weight of bumblebees vary very wide. Need body weight range.
- Disagree with removing body weight – there needs to be a discussion about what body weight data is being used for before agreeing to delete it from the test guideline. In theory by including bodyweight it allows expressing of toxicity relative to bodyweight to account for differences in sizes between different bee species. If in practice bodyweight is not being recorded for individual bees in accordance with the test guideline it should be discussed how bodyweight data is being used

3. para 33 - Delete the reporting of the weight of each bumblebee.



Agreement justification:

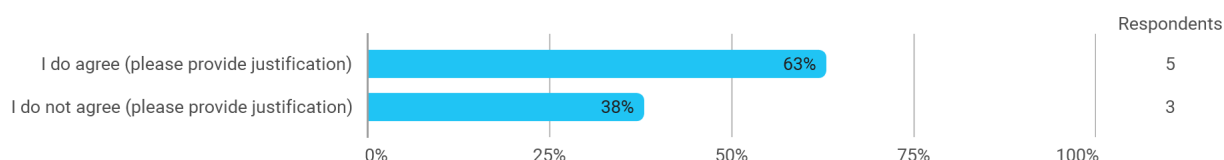
- Agreed. This action takes a lot of time and has no relevance for the study outcome or does influence the results in any way.

► Same as Suggestion 2

Disagreement justification:

- Unlike honey bees, bumble bees can vary widely in weight; having such data provides a means of assessing the extent to which weights actually varies
- See comment for suggestion 2
- We disagree. There needs to be a discussion about what body weight data is being used for before agreeing to delete it from the test guideline. If in practice bodyweight is not being recorded for individual bees in accordance with the test guideline it should be discussed how bodyweight data is being used.
- We disagree. Example should be given when this can be done as laboratories not familiar with this test may not know this.
- Disagree with removing body weight – there needs to be a discussion about what body weight data is being used for before agreeing to delete it from the test guideline. In theory by including bodyweight it allows expressing of toxicity relative to bodyweight to account for differences in sizes between different bee species. If in practice bodyweight is not being recorded for individual bees in accordance with the test guideline it should be discussed how bodyweight data is being used

4. para 19 - Delete Once during the experimental phase at least



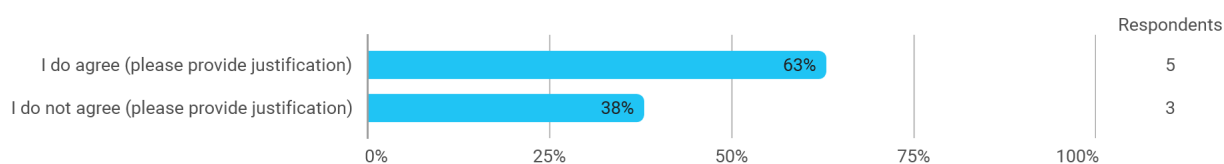
Agreement justification:

- only one application
- There is only one application

Disagreement justification:

- While it is implied, it may also need to be specified that analytical verification is needed on what is actually provided to the bees.
- Disagreed. Example should be given when this can be done as laboratories not familiar with this test may not know this.

5. para 20 - Delete whole paragraph



Agreement justification:

- not necessary

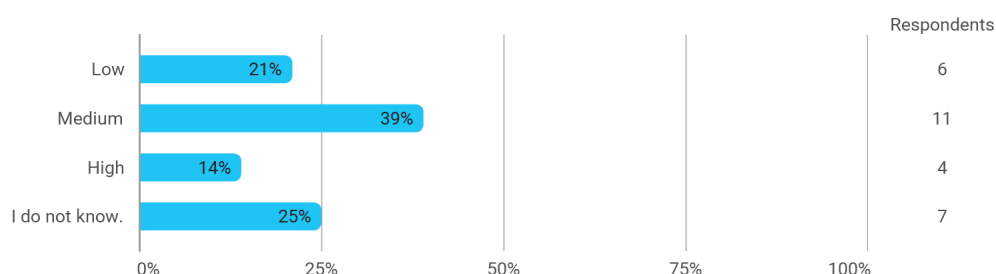
- This paragraph is not applicable

Disagreement justification:

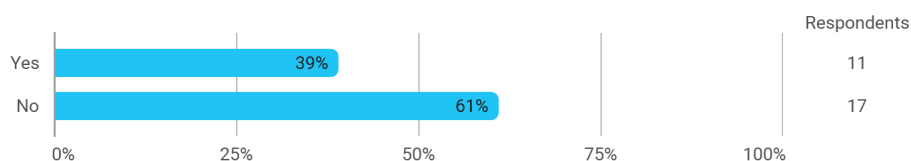
- While this is an acute toxicity test, if for any reason a new solution stock solution must be made up, then analytical verification of exposure is critical
- Disagreed to entirely delete but maybe to adapt. Proposal: If a new batch of the test chemical needs to be used during the test phase, a new Certificate of Analysis (re-analysis) should be provided in order to demonstrate that there are no major changes/deviations of the test item. Ideally studies should be conducted with the same chemical batch.
- We disagree. An alternative proposal could be: If a new batch of the test chemical needs to be used during the test phase, a new Certificate of Analysis (re-analysis) should be provided in order to demonstrate that there are no major changes/deviations of the test item. Ideally studies should be conducted with the same chemical batch.

B.9 Test No. 247: Bumblebee, Acute Oral Toxicity Test

Relevance of revision

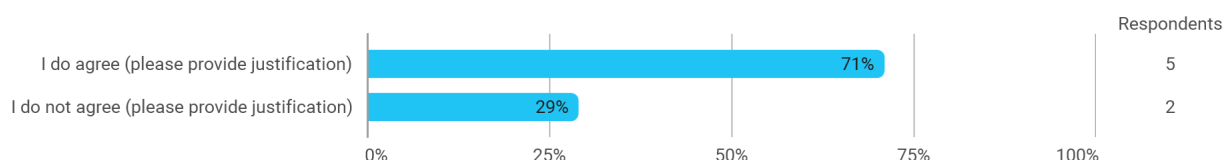


Interested to be part of an OECD Expert Group



The following suggestions have been collected:

1. para 21 - delete Once during the experimental phase at least...



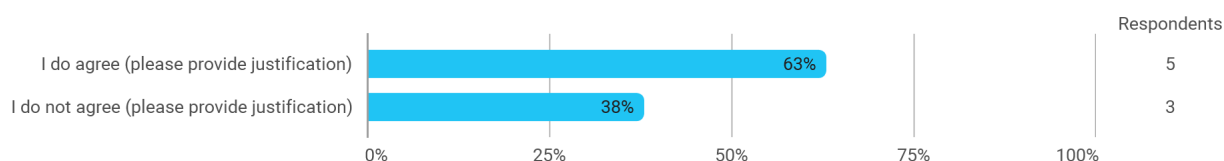
Agreement justification:

- only one application
- There is only one application

Disagreement justification:

- We disagree. Example should be given when this can be done as laboratories not familiar with this test may not know this.

2. para 22 - Delete whole paragraph



Agreement justification:

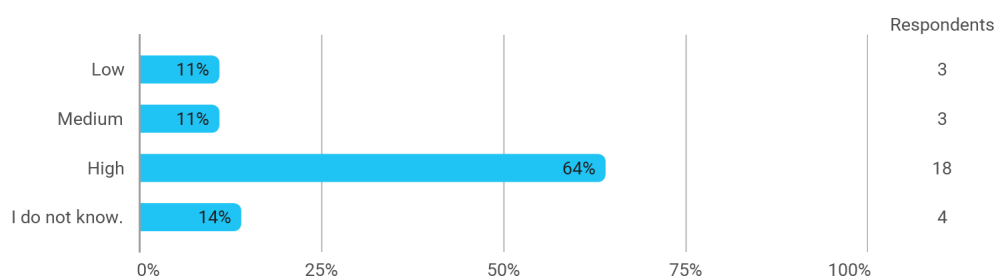
- not necessary whole paragraph
- There is only one application

Disagreement justification:

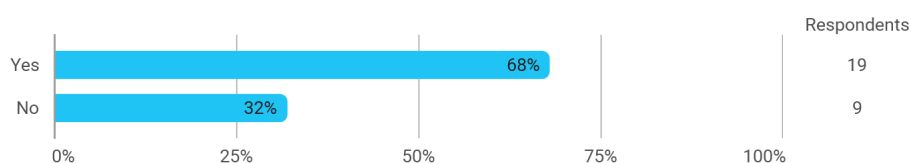
- While this is an acute toxicity test, if for whatever reason the stock solution may need to be changed, the exposure concentrations should be verified on what is actually administered.
- Disagreed to entirely delete but maybe to adapt. Proposal: If a new batch of the test chemical needs to be used during the test phase, a new Certificate of Analysis (re-analysis) should be provided in order to demonstrate that there are no major changes/deviations of the test item. Ideally studies should be conducted with the same chemical batch.
- We disagree. An alternative proposal could be. If a new batch of the test chemical needs to be used during the test phase, a new Certificate of Analysis (re-analysis) should be provided in order to demonstrate that there are no major changes/deviations of the test item. Ideally studies should be conducted with the same chemical batch.

B.10 Guidance Document 239: Honey Bee Larval

Relevance of revision

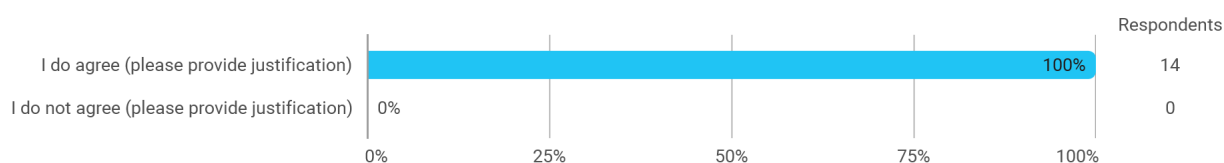


Interested to be part of an OECD Expert Group



The following suggestions have been collected:

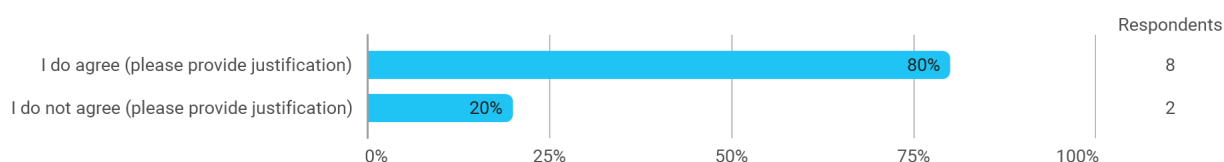
- para 7 - Test validity criteria 'in the control plate(s), the adult emergence rate on D22 should be $\geq 70\%$ across all replicates' - Suggested change – wording of validity criteria should be changed to make it clear that emergence on D22 should be $\geq 70\%$ as a mean across all replicates, not $\geq 70\%$ for every individual replicate Reason for change – Several industry companies have had 22 day larval studies rejected by Member States when the study does not meet $\geq 70\%$ for every individual control replicate . The ICPPR brood group, have contacted OECD to ask for clarifying statement to be added to the cover page of the Guidance 239. Discussions are ongoing, however it is hoped this is resolved relatively quickly, as it's causing many risk assessments not to be finalised.**



Agreement justification:

- ▶ 30% allowed emergence loss per replicate allows 3 bees not to hatch ($3 \times 3 = 9$); 30% of 36 bees allows 10 bees not to hatch
- ▶ Agreed. The revised language, currently on the cover page of the Guidance Document 239, should be moved to the body of the text.
- ▶ The way of interpretation of "across replicates" ("average" or "per replicate") has an immense influence on the validity of the test, therefore clarification is needed. It should be clearly stated, that "across replicates" means: %Emergence is calculated related to the total sum of individuals in all replicates, rather than per replicate.
- ▶ needed clearer validity criteria
- ▶ important to avoid misinterpretation
- ▶ Please be aware that there is an update available for OECD GD 239 (July 2021) where it states the following on the title page: "Following several questions raised after publication, the OECD would like to clarify that the validity criteria mentioned in paragraph 7 «across all replicates» should be understood as the average value across all replicates, a replicate being defined as the number of larvae originating from the same colony." The question remains, if this topic needs any additional revision as proposed here.
- ▶ already handled.
- ▶ A reasonable change that better accounts for variability in the test.

- para 6 - ... Technical grade or formulated dimethoate (CAS 60-51-5) or technical grade fenoxycarb**



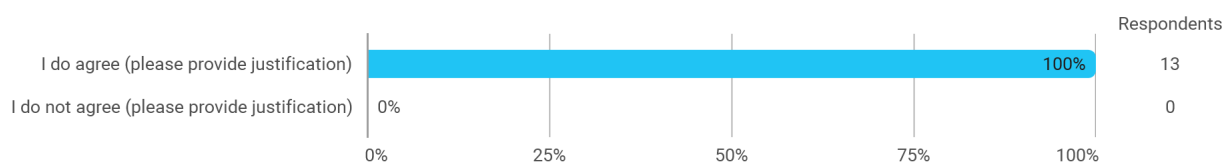
Agreement justification:

- ▶ Agreed. Formulated dimethoate is sufficient as a reference chemical.
- ▶ Agree – keep in mind that toxic reference items can be difficult to obtain so more than one option is needed for IGR and 'standard'
- ▶ We agree, no further justification needed
- ▶ Agree – keep in mind that toxic reference items can be difficult to obtain so more than one option is needed for IGR and 'standard'
- ▶ Need to have an alternative to dimethoate.

Disagreement justification:

- ▶ Both dimethoate and fenoxycarb can't be purchased legally in Germany right now. A different active would be needed. If TGAi or formulation is not relevant in my eyes.
- ▶ technical grade dimethoate is fine

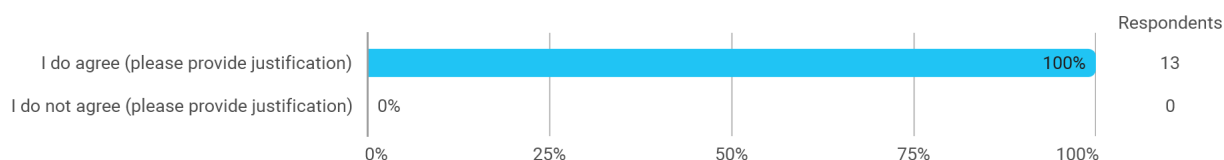
3. para 7 - In the control plate(s), cumulative larval mortality from D3 to D8 should be ≤ 15% in the average (mean) of all replicates.



Agreement justification:

- ▶ That was addressed in the latest update of the GD.
- ▶ Totally agree
- ▶ Agreed. Clarifying language to prevent misunderstanding by regulatory authorities
- ▶ same as for suggestion 1
- ▶ it is clear
- ▶ important to avoid misinterpretation
- ▶ Please be aware that there is an update available for OECD GD 239 (July 2021) where it states the following on the title page: "Following several questions raised after publication, the OECD would like to clarify that the validity criteria mentioned in paragraph 7 «across all replicates» should be understood as the average value across all replicates, a replicate being defined as the number of larvae originating from the same colony." The question remains, if this topic needs any additional revision as proposed here.
- ▶ A reasonable change that better accounts for variability in the test.

4. para 7 - In the control plate(s), the adult emergence rate on D22 should be $\geq 70\%$ in the average (mean) of all replicates.

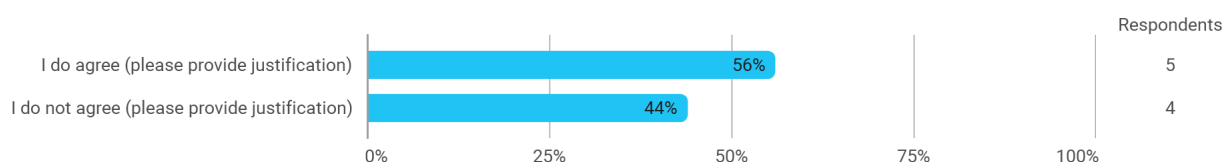


Agreement justification:

- ▶ That was addressed in the latest update of the GD.
- ▶ Agreed. Clarifying language to prevent misunderstanding by regulatory authorities
- ▶ same as for suggestion 1
- ▶ here also
- ▶ important to avoid misinterpretation
- ▶ Agree Please be aware that there is an update available for OECD GD 239 (July 2021) where it states the following on the title page: "Following several questions raised after publication, the OECD would like to clarify that the validity criteria mentioned in paragraph 7 «across all replicates» should be understood as the average value across all replicates, a replicate being defined as the number of larvae originating from the same colony." The question remains, if this topic needs any additional revision as proposed here
- ▶ A reasonable change that better accounts for variability in the test.

Disagreement justification:

5. para 10 - ... If the larvae already show signs of pre-pupation on day 7, the saturated potassium sulphate (K₂SO₄) solution can be replaced by a saturated sodium chloride (NaCl) solution on day 7 already instead of day 8 to achieve an adequate relative humidity during pupation phase.



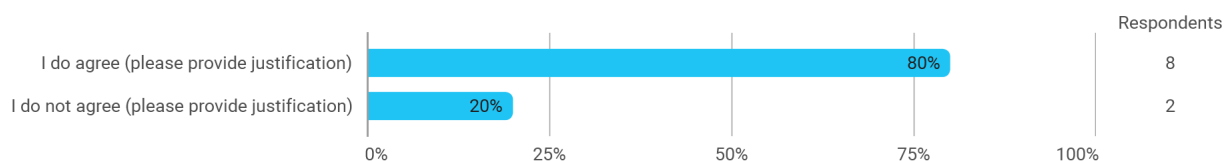
Agreement justification:

- ▶ If this is better to develop bees, I agree.

Disagreement justification:

- ▶ Using salt should be optional, because many contract labs use climate chambers with adjustable humidity. If the humidity is decreased on D7 or D8 is not essential in my opinion.
- ▶ Disagree - Option should be included to transfer pre-pupae to a separate plate. The "pupal plate" can then be placed in a desiccator with a dish filled with NaCl solution while larvae that have yet to reach pre-pupae can continue to develop in the presence of the K₂SO₄ solution.

6. para 11 - Alternatively, the plates can be closed with the lid and transferred into an incubator at 34-35°C with a relative humidity within the range 50 - 80%. On day 22 the fully developed, adult honey bees are counted as emerged.



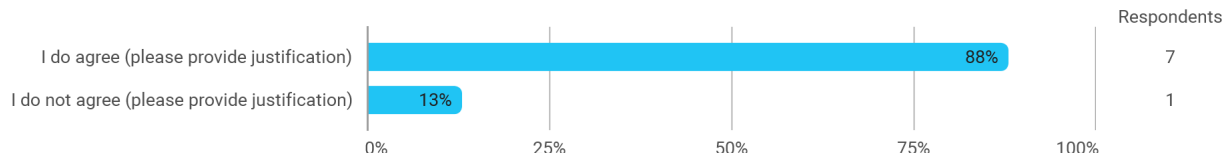
Agreement justification:

- ▶ Agreed. This is critical if weights of enclosed adults is taken, which is a request by certain regulatory authorities (i.e., USEPA)
- ▶ If with lid is better, I agree.
- ▶ We agree. No further justification needed.

Disagreement justification:

- ▶ by not feeding the freshly emerged bees they were killed not in a human manner
- ▶ It is more beneficial to animal wellbeing not to let the bees slowly starve to death. Also it is easier to assess, which bees are fully emerged.

7. para 31 - The plates of the dimethoate reference item treatment can be disposed already after the assessment on D8.



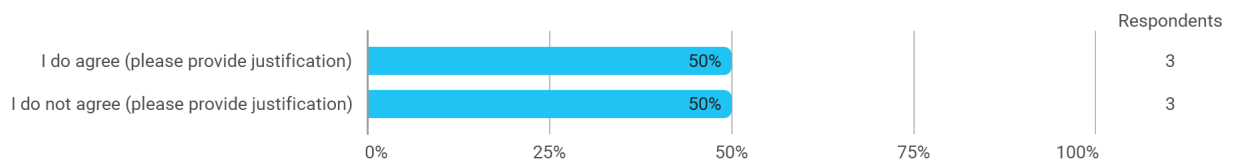
Agreement justification:

- ▶ dimethoate reference item affected during the larval stage
- ▶ We agree. However, we propose to add the following text: "If fenoxycarb is used, observations should continue until D22 (i.e., test termination)".

Disagreement justification:

- ▶ Disagreed. If this modification is included in the revised guidance, there should be explicit differentiation between dimethoate and fenoxycarb when used as reference items. For example, "If dimethoate is used, the plates of the dimethoate reference item treatment can be disposed after the assessment on D8. If fenoxycarb is used, observations should continue until D22 (i.e., test termination)".

8. para 35 - adverse effects: larval mortalities from D3 to D8, pupal mortalities from D8 to D22 and emergence rate on D22. ...



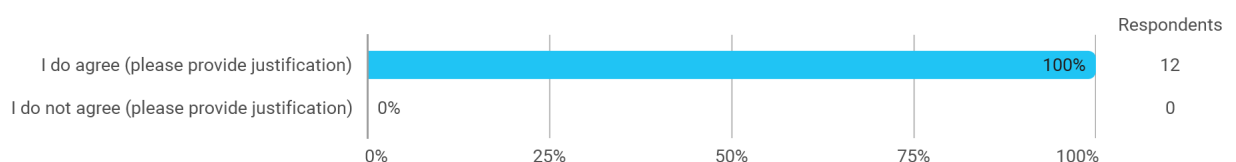
Agreement justification:

- pupal stage is D8 to D22

Disagreement justification:

- The named endpoints should be listed in a table, but it seems sufficient to perform the full statistical evaluation just for D22.
- Disagreed. All changes are ok, however D8 is repeated for both larval mortalities and pupal mortalities. It should be changed to D9-D22 for pupal mortalities.
- We disagree. As the emergence rate on D22 is the inverse % of mortality we do not see the need to add an additional mortality percentage on day 22. In addition, D8 is repeated for both larval mortalities and pupal mortalities. Pupal mortality assessments should be start from D9 and onwards.

9. para 35-37 data and statistical analysis – it is suggested that clarity is given on what constitutes a replicate for statistical analysis. Larvae are collected from three different colonies each representing a replicate. Paragraph 22 states ‘a total of seven to eight (if solvent is used) well-plates are used per test. Each group of a minimum of 12 larvae from each of the three colonies is considered a replicate for a given treatment level ad identified as such on the microplate’. Some regulatory authorities consider each individual cell (well-plate) containing a larva a replicate therefore the guidance document would benefit from additional wording to clarify



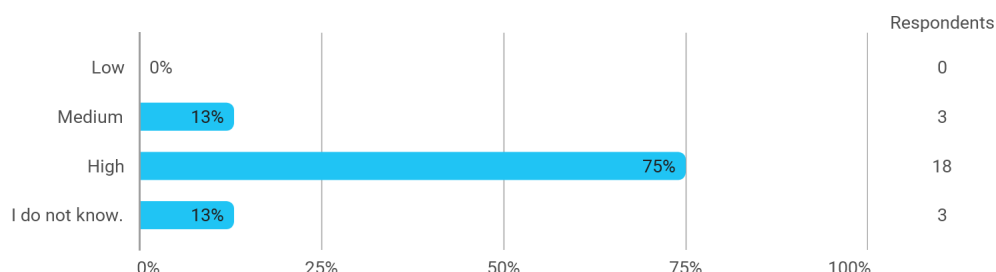
Agreement justification:

- each colony represents a replicate - in this way it is possible to evaluate colony effects
- A replicate should be defined (as it momentarily is) as 12 bees of the same colony, because they are all originating from the same queen.
- Totally
- Agreed. Each colony represents a replicate for the purposes of statistical analyses.
- additional wording is better to understand
- clarification is needed

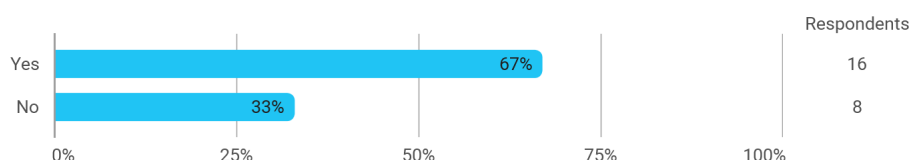
- ▶ Agree – i.e. EPA using each well as a replicate which gives large number of replicates in comparison to considering only the colonies a replicate – in which case you only 3 true replicates (the rest are pseudo replicates)
- ▶ We agree, no further justification needed.
- ▶ The TG would benefit as defining each plate I(*not each well in the plate) as the replicate.

B.11 Test No. 208: Terrestrial Plant Test: Seedling Emergence and Seedling Growth Test

Relevance of revision

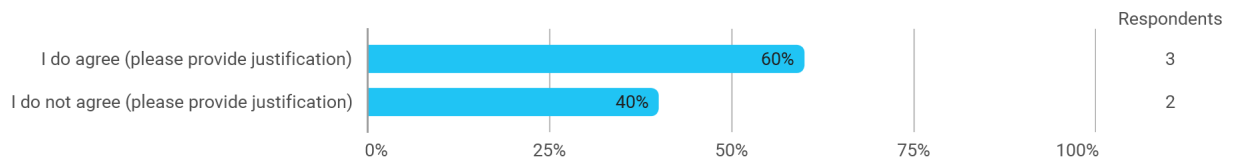


Interested to be part of an OECD Expert Group



The following suggestions have been collected:

1. **para 2- The test assesses effects on seedling emergence and early growth of higher plants following exposure to the test substance in the soil (or other suitable soil matrix). Seeds are placed in contact with soil treated with the test substance and evaluated for effects following usually 14 to 21 days after 50 % emergence of the seedlings in the control group. Endpoints measured are visual assessment of seedling emergence, dry shoot weight (alternatively fresh shoot weight) and in certain cases shoot height, as well as a qualitative or semi-quantitative assessment of visible detrimental effects on different parts of the plant. These measurements and observations are compared to those of untreated control plants.**
Rational: Visual injury assessments are no measurements, they are a visual scoring/rating of the level of injury compared to some (visually chosen) representative control plants. Although the scoring of visual symptoms is highly dependent on personal expertise and perception, the scoring scale can be adjusted within a laboratory through regular training session and peer comparison. However, there may be substantial variation between different laboratories and therefore endpoints based on visual injury would not be comparable between labs. This hampers the reliability of such qualitative endpoints compared to measured endpoints.



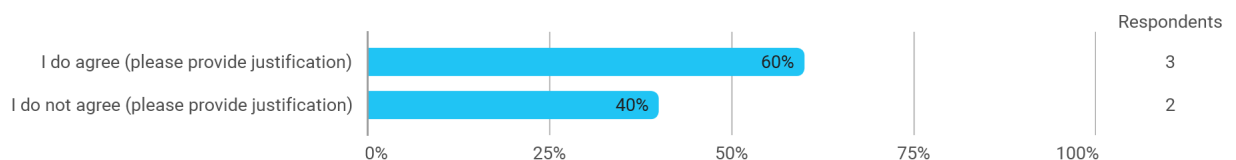
Disagreement justification:

- ▶ assessments at 7, 14 and 21 days after 50% emergence in the control; survival is missing; emergence as visual assessment (living and dead seedlings are counted (quantal data)); spray application or incorporation into the soil
- ▶ (i) The sentence lists the endpoints (i.e. variables) measured. No further statement on the way of measurement is needed in this paragraph “principle of the test”. Details of the measurement should be given under “...” and “reporting”. (ii) A quantitative assessment of visible detrimental effects is preferenciated, whenever possible.

2. “Endpoints measured are dry shoot weight...” Please change to “Variables measured are dry shoot weight ...”

Rational: In ecotoxicology, an endpoint contains the information about the estimate used, effect threshold as well as biological variable.

This means that an endpoint can be derived for several variables (i.e., in the current TG e.g. ER50 for germination, plant height, biomass ...)



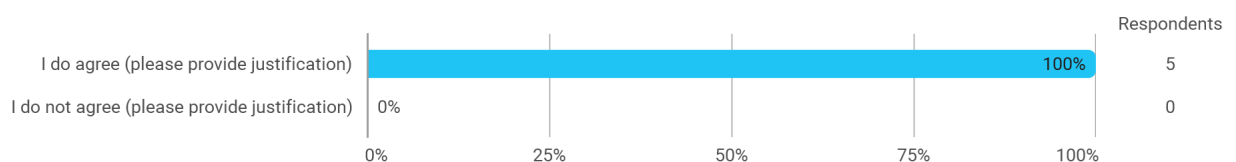
Agreement justification:

- ▶ Update suggested reflects more recent state of the art approach..
- ▶ Disambiguous

Disagreement justification:

- ▶ variable measured only at the end of the test
- ▶ Endpoints are correct

3. para 4 - “Also, the no observed effect concentration (NOEC), resp. no observed effect rate (NOER), and lowest observed effect concentration (LOEC), resp. lowest observed effect rate (LOER), can be calculated in this test.”

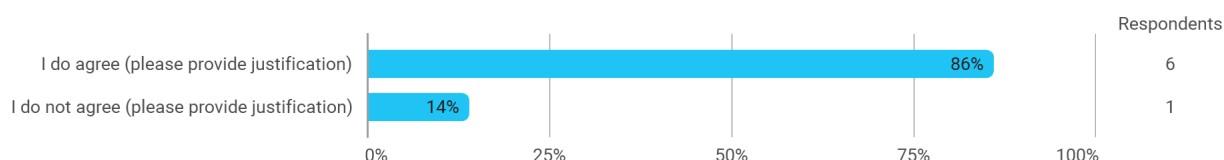


Agreement justification:

- ▶ clarification
- ▶ Test item units are mostly given as rate, e.g. g/ha.

- ▶ the correct dimension in terrestrial plant studies is rate/ha not concentration, hence clarification of endpoints a NOER, LOER, etc. is reasonable
4. **para 4 - The test can be conducted in order to determine the dose-response curve, or at a single concentration/rate as a limit test according to the aim of the study. If results from the single concentration/rate test exceed a certain toxicity level (e.g., whether effects greater than x% are observed), a range-finding test is carried out to determine upper and lower limits for toxicity followed by a multiple concentration/rate test to generate a dose-response curve. For quantitatively measured endpoints, an appropriate statistical analysis is used to obtain effective concentration ECx or effective application rate ERx (e.g., EC25, ER25, EC50, ER50) for the most sensitive parameter(s) of interest. Also, the no observed effect concentration (NOEC) and lowest observed effect concentration (LOEC) can be calculated in this test. Some of the data generated in phytotoxicity tests are qualitative, such as ratings based upon visual symptoms of phytotoxicity. Qualitative data such as ratings are not statistically analysed for these tests but may be used to report qualitative no-effect levels.**

Rational: This is in line with international guidelines, cf. OCPPS Guidance 850.4000, p.12, "(2) Calculation of endpoints (i) Backgrounds (A) qualitative data. Some of the data generated in phytotoxicity tests are qualitative, such as ratings based upon visual symptoms of phytotoxicity. Qualitative data such as ratings are not statistically analyzed for these tests but may be used to report qualitative no-effect levels."



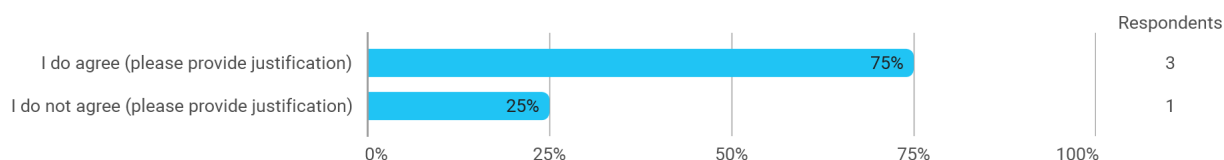
Agreement justification:

- ▶ no stat. analysis of a subjective measurement like Visual Injury
- ▶ NOER for phytotoxicity is okay, an ERx calculation is nonsensical
- ▶ quantitative evaluations stays subjective, not measurable
- ▶ effects greater x% - definition would help to decide yes or no
- ▶ if quantitatively measured data is available it does not make much sense to use qualitative data based on variable visual estimations for statistical calculations. In addition, the approach of not using qualitative data for ERx calculation is in line with other international Guidance (OCPPS Guidance 850.4000).

Disagreement justification:

- ▶ The observed visual effects are to be quantified - where possible - and an ERx derived. Please see also point 16.
5. **para 4 - Appropriate statistical analysis is used to obtain effective concentration ECx or effective application rate ERx (e.g., EC25, ER25, EC50, ER50) for the most sensitive parameter(s) of interest and change "parameter(s)" to "variable(s)"**

Rational: Variables are measured, parameters are fixed in the experimental design, e.g. light, temperature, test duration, treatments are parameters.



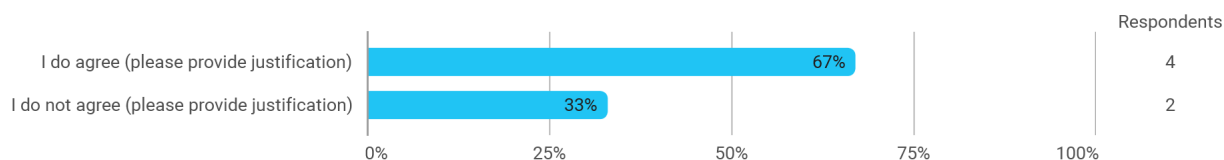
Agreement justification:

- Update suggested reflects more recent state of the art approach.

6. **para 6 - The validity criterion in the first bullet point of §6 is not clear. It says: “In order for the test to be considered valid, the following performance criteria must be met in the controls: the seedling emergence is at least 70%; ...” Does this mean that each individual replicate should have at least 70% emergence, or does it mean that the mean emergence rate across replicates for the control group should be at least 70%? This validity criterion should be clarified.**

Rational: This question was raised by the OECD secretariat to the NC for the ENV in an e-mail 2021-05-05. Based on the feedback received, OECD made the following conclusion (e-mail 2021-06-18):

“From feedback received it appears that the validity criterion in TG 208 is mostly interpreted as applying to the mean of all replicates in the control group, rather than to each replicate (in TG 208 a replicate can contain one or two large seed(s) only, depending on the size of the seed). For now, I will respond in that sense to the enquiry received. However, your feedback also indicates that it might be welcome to report the variance across replicates. The TG 208 is generally not very prescriptive for the test design (e.g. no minimum number of replicates), nor does the test report explicitly requires individual replicate data it seems. Like for many TGs drafted 10-15 years ago or more, the level of prescription is not high. If the TG 208 was to be written today, more precision would be expected, based on experimental validation work. I also take from feedback received that it may not be appropriate to have the same way of interpreting the validity criterion in all Test Guidelines, as test designs vary. Thank you and for now, I will leave this conversation here. If there was a need to re-open and clarify things further, a clear project proposal would be needed.”



Agreement justification:

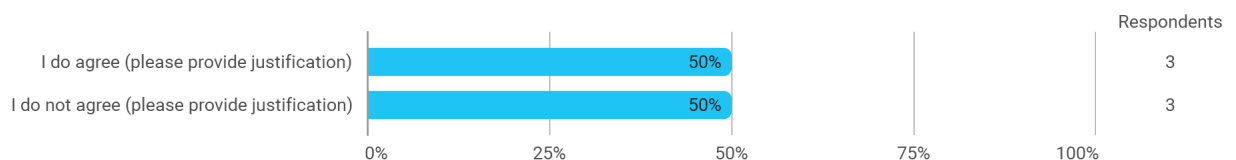
- The way of interpretation of "across replicates" ("average" or "per replicate") has an immense influence on the validity of the test, therefore clarification is needed. It should be clearly stated, that “across replicates” means: %Emergence is calculated related to the total sum of individuals in all replicates, rather than per replicate.

- ▶ preselected answer does not fit! In general such a validity criterion is defined as average across all control replicates (cf. OECD GD 239 Honey bee larva, repeated exposure). OECD could add a clarification statement as done in OECD GD 239

Disagreement justification:

- ▶ should be always be on treatment level not rep
- ▶ No clarification is needed to perform a risk assessment since the main criterion is plausibility of the results.

7. para 6- The validity criterion regarding the emergence in control pots should be clarified. Does it refer to each individual control or to the average?



Agreement justification:

- ▶ The way of interpretation of "across replicates" ("average" or "per replicate") has an immense influence on the validity of the test, therefore clarification is needed. It should be clearly stated, that "across replicates" means: %Emergence is calculated related to the total sum of individuals in all replicates, rather than per replicate.

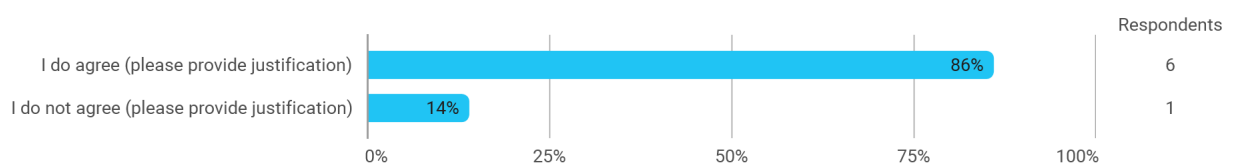
Disagreement justification:

- ▶ per test species!?
- ▶ No clarification is needed to perform a risk assessment since the main criterion is plausibility of the results.

8. para 7 -Deletion of Reference Substance Chapter.

Rational:

- - no specification of reference item and effect available that would need to be achieved
- - each species and test system (Seedling Emergence vs. Vegetative Vigour) might need a different reference substance
- - detecting no resistance to a certain Mode of Action does not mean that resistance to another MoA (e.g. that of the planned test items) could automatically be excluded



Agreement justification:

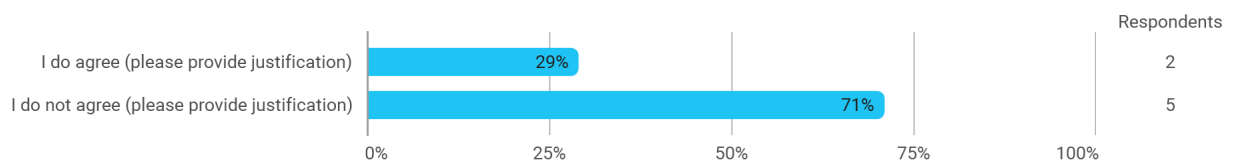
- ▶ no tests with Reference Substances
- ▶ not needed, internal control

- ▶ is not relevant,
- ▶ has not been used in past. potentially several toxic reference substances might be necessary (dicots/monocots). Potentially a different reference might need to be applied for other types of plant studies.

Disagreement justification:

- ▶ no conclusion yet, justification: the issue might be discussed (see point 9.)

9. para 7 - Some suitable reference substances should be added: e.g., sodium trichloroacetate or boric acid.



Agreement justification:

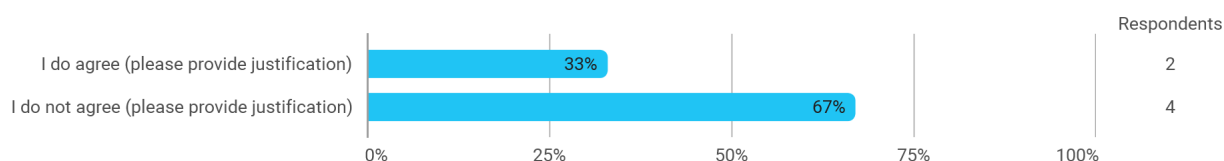
- ▶ no conclusion yet justification: more information needed to start a discussion

Disagreement justification:

- ▶ no tests with Reference Substances at all
- ▶ No reference substance necessary
- ▶ not needed, internal control
- ▶ is not relevant
- ▶ unless ring-tested no toxic reference should be made mandatory

10. para 19 - The number of plants per pot will depend upon the species, pot size and test duration, and should provide adequate and uniform growth conditions and avoid overcrowding and root partitioning of plants by each other for the duration of the test. [...] As an example, one to two corn, soybean, tomato, cucumber, or sugar beet plants per 15 cm container; three rape or pea plants per 15 cm container; and 5 to 10 onion, wheat, or other small seeds per 15 cm container are recommended. Any deviations from these recommendations should be justified. Plant growth should be documented with pictures. Statistical analysis should demonstrate that any deviation does not impact the variability of the measured variables (i.e., increase the variability; see validity criteria). The number of seeds and replicate pots....

Rational: This is the most violated recommendation of the TG. Indeed, most of the studies submitted are performed at higher plant density, without any rational about the potential impact of the deviation on the results. However, there are some indications in the literature that plant density directly impacts plant growth. Poster Teresiak, SETAC 2015; poster Simoneit-Gast, SETAC GLB 2015.



Agreement justification:

- Plant density has an impact on plant growth and thus might affect the results of the study. (refers to points 10-13)

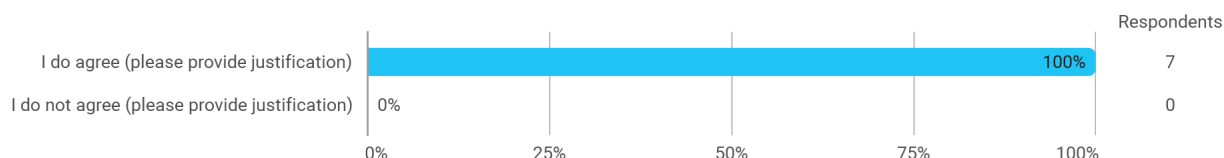
Disagreement justification:

- is it limited to pot with 15 cm diameter? no. of seeds for example: 2-3 oilseed rape, 4-6 onion, wheat, oat...; to minimize the variability may be 2 to 3 pots per replicate should be used (see OCSPP 850.4100)
- the effort of making pictures of for all plant species would be enormous
- plant density has to be discussed but verification of deviations difficult
- this point is a recommendation not a fixed requirements. While for agrochemicals the number of plants/pot have been reduced in recent years, the original recommendations may still be sufficient for studies intended for REACH or Biocide submissions. It would be more useful to clarify conditions for the definition of plant number per pot. This would also serve as guidance when new species are included in studies. Documentation with pictures can be useful, but it should not become a strict requirement (pictures of what exactly, how many, all individuals vs. examples, timings?) It's unclear what parameters of plant growth are meant to be evaluated in statistical analysis.

11. para 19 - "The maximum plant density would be around 3 - 10 seeds per 100 cm² depending to the size of the seeds. As an example, one to two corn, soybean, tomato, cucumber, or sugar beet plants per 15 cm container; three rape or pea plants per 15 cm container; and 5 to 10 onion, wheat, or other small seeds per 15 cm container are recommended."

Please clearly specify the pot size in the examples (i.e., the diameter for a round pot and the side of a square pot). Alternatively, specify the minimum area per seed recommended for each species referred to in the example.

Rational: Somewhat clearer guidance on this point might lead to less discussions when evaluating the tests.

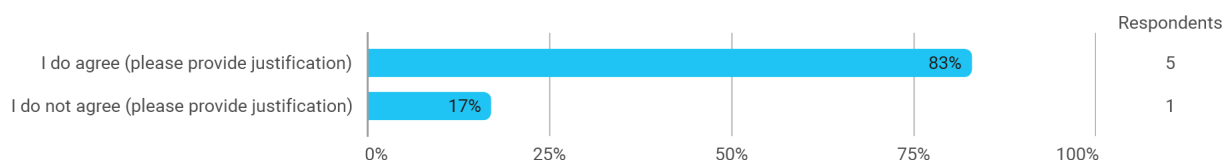


Agreement justification:

- Please clearly specify
- has to be clear
- discussion needed

- ▶ Clarification needed
- ▶ Plant density has an impact on plant growth and thus might affect the results of the study. (refers to points 10-13)
- ▶ it is more pragmatic to define plant number/pot by area required/species. Also in line with suggestion #12 nutrient requirements and external nutrient supply are highly important. sufficient nutrient supply may compensate for reduced area/plant available as growth is more dependent on nutrients than on space (at least during the first 3 weeks of growth)

12. para 19 - Number of seeds per pot. In quite some tests the number of seeds per pot is higher than recommended in the test guideline. Despite this, the plants are performing well and the test can be considered valid. Maybe the recommendation regarding the number of seeds per pot should be adjusted in a certain way.



Agreement justification:

- ▶ Please clearly specify
- ▶ see 11
- ▶ Clarification needed
- ▶ see #11

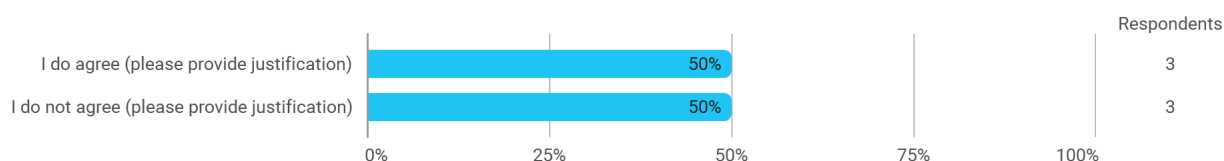
Disagreement justification:

- ▶ Plant density has an impact on plant growth and thus might affect the results of the study. (refers to points 10-13)

13. para 19 - The maximum plant density should be 2 - 5 seeds per container (based on a diameter of 15 cm), based on plant size. As an example, 2 seeds should be used for all dicotyledonous species and maize (Zea maize) and 4 seeds for all monocotyledonous species except of maize, based on a container with a diameter of 15 cm.

Rational: - Currently, the stated range for test conditions is narrow and based on what can technically be achieved in a growth chamber which is a highly controlled environment. The test conditions should rather reflect the (wider) range of conditions for optimal plant growth which can also be different for different plant species. In addition, tests are mainly conducted under greenhouse conditions which can be less climatically controlled than growth chambers, e.g. light intensity can easily be higher than $350 \pm 50 \mu\text{E}/\text{m}^2/\text{s}$ during summer months.

- Short deviations from the specified range (e.g. temperature) can often not be avoided despite efforts and good equipment. These deviations do not have an impact on the study quality since comparison is made to the control group under the same conditions. However, it is often a point of criticism by authorities.



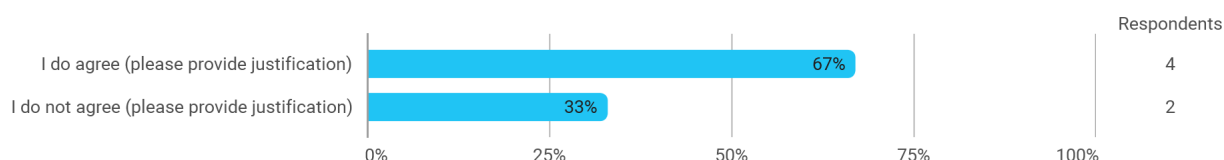
Agreement justification:

- Clarification needed
- no conclusion yet justification: Agree that information on pot size is insufficient and inconsistent. However, the proposal might need to be discussed. Plant density has an impact on plant growth and thus might affect the results of the study. (refers to points 10-13)

Disagreement justification:

- insert only as an example and recommendation, test species are missing...
- number of seeds per pot should be adjusted to the size the species can reach within the standard duration of 3 weeks. the nutrient requirements and whether these can be satisfied by external supply also need considerations

14. para 22 - Amendment of the ranges of test conditions to better reflect requirements for adequate plant growth / Inclusion of a statement which acknowledges the difficulty to maintain the exact conditions of the guideline (also depending on greenhouse equipment, season etc.) and allows for deviations over a short period of time (x number of hours).



Agreement justification:

- would prevent discussions
- pragmatic, greenhouses are usually not 100% controlled; temperature and sunshine are often buffered with certain delay, extreme outside conditions cannot be buffered completely (technical issues)

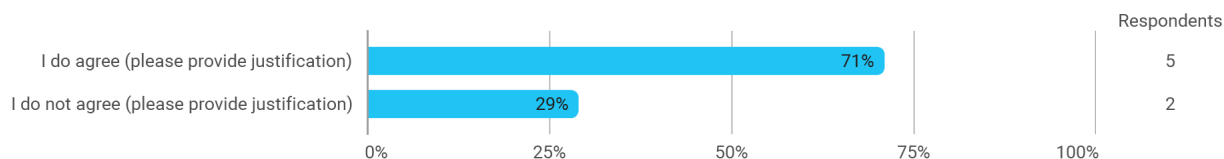
Disagreement justification:

- Glasshouse test are susceptible to local weather conditions. The issue of variation in weather conditions over time and between locations is best addressed via observation of control plant performance and good plant husbandry.
- A more detailed proposal is appreciated for the discussion of this item. Strictness of test conditions should be maintained, as also short-time deviations may be harmful for plant growth.

15. para 22 - Light intensity should be > 250 $\mu\text{mol}/\text{m}^2/\text{s}$; If intensity decreases below 250 $\mu\text{mol}/\text{m}^2/\text{s}$, additional lighting is required.

Rational: The use of a phytotoxicity rating system assessing visible herbicide injury in percent affected tissue (e.g. non-green tissue) rather than an injury rating (e.g. rating

1–5, for no, slight, moderate, severe, total plant symptoms, respectively) avoids having to deal with analyzing categorical data. Given a dose-response-relationship ERx for visual effects compared to un-exposed controls, as required by regulatory authorities, could more easily be derived. Additionally, using the uniform scale of “percent effect” for all endpoints the effects seen are more easily comparable.



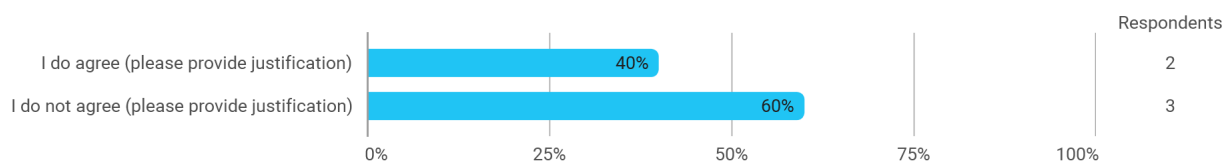
Agreement justification:

- ▶ light intensity > 400 μmol cannot be prevented and are normal in a greenhouse
- ▶ the actual light intensity is not needed for perfect growing conditions
- ▶ no conclusion yet justification: General agreement to have more precise information on light intensity. More information is appreciated for the discussion of this item.
- ▶ Reasonable

Disagreement justification:

- ▶ good experience with 200 $\mu\text{mol}/\text{m}^2/\text{s}$
- ▶ Currently min light intensity is 300 $\mu\text{mol}/\text{m}^2/\text{s}$

16. para 29 - "A uniform scoring system for visual injury should be used to evaluate the observable toxic responses. Examples for performing qualitative and quantitative visual ratings are provided in references (23) (24).As indicated there, preferably quantitative data (e.g. from 0 to 100 %) should be used to allow ERx calculation when multiple rate testing has been conducted to generate a dose-response curve."



Agreement justification:

- ▶ Please see also point 4. The observed visual effects are to be quantified - where possible - and ERx derived.

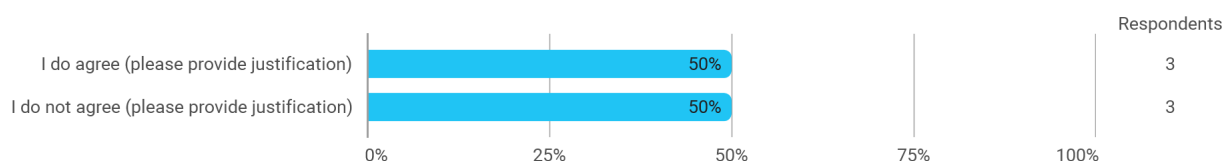
Disagreement justification:

- ▶ no ERx calculation for phytotoxicity. Examples only - leave room for the study plan (very subjective evaluation, difficult to compare).
- ▶ I still think a visual rating is not to be statistically analysed
- ▶ visual injury is a qualitative parameter mainly needed for the validity criterion of "no phytotoxicity in the control" and to ensure that late onset of injury is detected (->prolongation). Visual estimations of injury are variable between individual assessors and

along the scoring scale (larger in the middle range). The Guideline already includes several quantitatively measured parameter, hence there is no need to squeeze a qualitative parameter artificially into quasi-quantitative frame.

17. para 29 - Temperature for determination of constant weight of plant shoots should be given as a range, i.e. 60 ± 5 °C

Rational: It is common practice in the analysis of continuous biomass data (i.e. shoot dry and fresh weight as well as shoot height) that they are rescaled by calculating the effect size in percent. Rescaling of data is considered as convenient because two-parametric dose-response models can be used for dose-response analysis. However, this is not state of the art anymore and is also not in line with the OCED GD 54 “Current Approaches in the Statistical Analysis of Ecotoxicity Data”, since rescaling reduces the variation of the data by excluding the natural variability (i.e. control values are missing in dose-response analysis). Moreover, two-parametric models generally have a poorer fit to continuous data compared to three-parametric models, which affects the precision and accuracy of derived ERx values. Therefore, any ad hoc adjustments of continuous data before fitting a dose-response model is inappropriate and should be avoided.



Agreement justification:

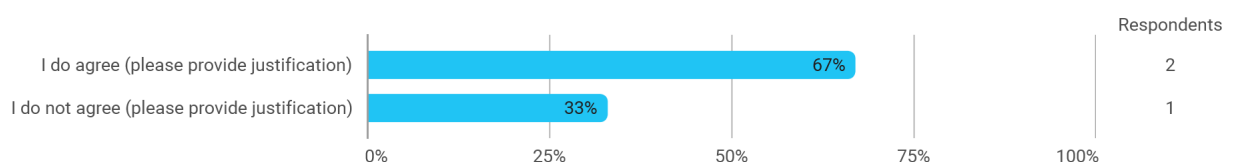
- ▶ when is constant weight reached? E.g. deviation <1% after 3 hours
- ▶ is it needed to settle a temperature at all? the main think is constant weight which has to be demonstrated

Disagreement justification:

- ▶ is a range 20 ± 10 °c
- ▶ Visual effects invariably translate into effects on biomass and shoot length. These parameters are already assessed and provide a qualitative, objective assessment of effects. In contrast, visual scores are subjective, lack granularity and do not provide a measure of variability. It is also uncertain whether estimation of endpoints based on visual observations will add value to risk assessment given that this TG already generates several endpoints. The issue of visual score endpoints is currently being evaluated by a WG working under SETAC Plant Interest Group.
- ▶ no conclusion yet justification: A range might be reasonable. However, the range proposed for drying temperature is too high to allow standardisation. The content of water remaining in the dried plant tissue impacts the biomass.

18. para 31 - “For the growth of the seedlings (weight and height) as continuous endpoints ECx or ERx and its confidence limits can be estimated by using appropriate regression analysis (e.g. Bruce-Versteeg non-linear regression analysis (25)) by using three-parametric models, which includes the control as additional Parameter.”

Rational: Dose-response models used to derive ERx values for biomass measurements (i.e. shoot dry and fresh weight as well as shoot height) typically encompass effect sizes from 0% to 100%. In terms of an 100% effect in a treatment compared to control, this means that there is zero biomass in that treatment, which can only be achieved if the plants are fully mineralized. As this is impossible due to the study design, the closest approximation to an effect size of 100% is given by setting the plants to zero that died or did not emerged during the experiment. However, this is not explicitly mentioned in the OECD GD 208, which results in statistical evaluation of biomass effects in which dead or not emerged individuals are excluded from dose-response analyzes. This, in turn, violates statistical rules and subsequently leads to false endpoints for environmental risk assessment, which constitute a serious problem and therefore should be corrected as soon as possible by including the suggested sentence above.



Agreement justification:

- Update suggested reflects state of the art.

Disagreement justification:

- think the existing description is sufficient

19. para 31 - “For the growth of the seedlings (weight and height) as continuous endpoints ECx or ERx and its confidence limits can be estimated by using appropriate regression analysis (e.g. Bruce-Versteeg non-linear regression analysis (25)). Note that the biomass of individuals that died during the study should be set to zero and included in the dose-response analysis.”

Rational: As mentioned under 2., an endpoint can be calculated for different biological variables. Therefore, it is important that ERx values are reported for each variable measured in the test. This includes visual injuries. This is supported by the fact that visual injury (i.e., phytotoxicity) is clearly mentioned in the TG as relevant variable: i) in the principle of the test as being a relevant endpoint: “Endpoints measured are visual assessment of seedling emergence, dry shoot weight (alternatively fresh shoot weight) and in certain cases shoot height, as well as an assessment of visible detrimental effects on different parts of the plant” (p. 1).

ii) in the ERx definition in Annex 1: “ECx. x% Effect Concentration or ERx. x% Effect Rate is the concentration or the rate that results in an undesirable change or alteration of x% in the test endpoint being measured relative to the control (e.g., 25% or 50% reduction in seedling emergence, shoot weight, final number of plants present, or increase in visual injury would constitute an EC25/ER25 or EC50/ER50 respectively).” Although plant visual injury is commonly measured and reported in almost all studies, this variable is not always assessed in a quantitative way, i.e. ERx values are rarely derived. In study reports the assessment of this variable is often limited to a quantitative description of the effects observed. However, for the risk

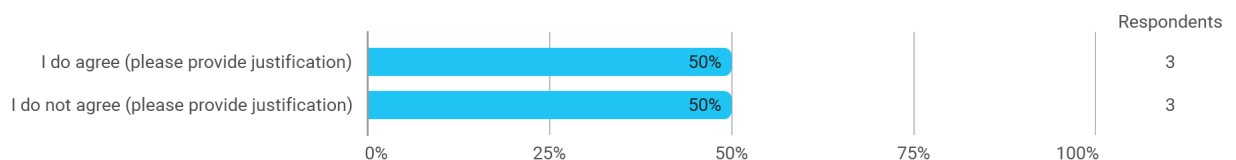
assessment of plant protection products an agreement has recently been met at EU level and between all Member States of the European central Zone to consider this effect in a quantitative way in the risk assessment (see references below). Therefore, it is of most importance that ERx value are now calculated. The report of the normalized confidence limits (i.e., confidence limits / the corresponding ERx values) is more informative.

Publications:

EFSA Supporting publication 2019:EN-1673

Central Zone Evaluation Manual, a Working document on Risk Assessment of Plant Protection Products in the Central Zone - Ecotoxicology

Version 1.0, May 2021.



Agreement justification:

- ▶ as well plant height of individuals that died during the study should be set to zero and included in the dose-response analysis
- ▶ Plants cannot be defined as dead (roots are not assessed). Therefore, assessment parameters should only be set to zero when there is an absence of measurable biomass.

Disagreement justification:

- ▶ dead individuals are addressed by the endpoint survival and should not be included in other evaluations like dry weight determination
- ▶ Disagree to the sentence “Note that the biomass of individuals that died during the study should be set to zero and included in the dose-response analysis”(1) Including dead individuals in evaluation of weight seems to be questionable at all, since evaluation of mortality should not be mixed up with evaluation of growth. (2) How to define “dead” in the context of plants? (3) A metric variable such as weight or height can never be zero. If there is anything measurable, then there is a number available. If there is nothing measurable, then it is a missing value – not zero. Zero is an arbitrary assumption, not a measurement result. Setting metric data to zero will result in a significant reduction of the treatment mean. Thereby, mortality is mixed up with growth, hence, results of different studies are no longer comparable.
- ▶ no conclusion yet justification: General agreement that excluding the most severely affected individuals from ERx determination biases the results. However, following that proposal different effect parameters are mixed up. A discussion on this item is appreciated.

20. para 32 - percent visual injury and qualitative and quantitative description of visual injury.

Replace by:

- a qualitative description and a quantitative assessment of the visual injuries, i.e., ERx calculation, (chlorosis, ...
- description of the rating scale used to judge visual injury, if visual rating is provided;

Replace by:

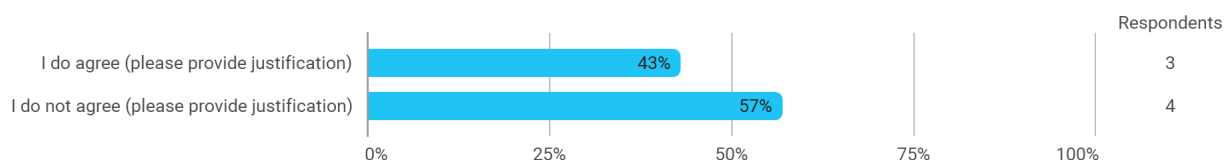
- description of the rating scale used to judge visual injury, that provides the information needed to verify the quantitative assessment of the phytotoxicity data (e.g., ERx calculation);

- ECx or ERx (e.g. EC50, ER50, EC25, ER25) values and related confidence limits.

Replace by:

- ECx or ERx (e.g. EC50, ER50, EC25, ER25) values and related normalised confidence limits.

Rational: Not reporting the storage conditions of the seeds should not be a reason to always reject the study.



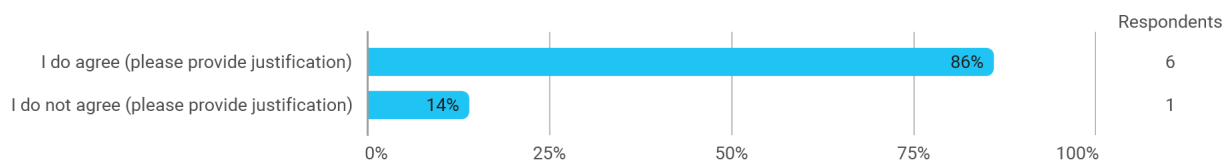
Agreement justification:

- ▶ the handling of dead plants is often point of discussion
- ▶ The information is necessary to evaluate the test.

Disagreement justification:

- ▶ stat. analysis is not reliable for visual injury observations
- ▶ no ERx calculation for phytotoxicity
- ▶ The issue of visual score endpoints is currently being evaluated by a WG working under SETAC Plant Interest Group.
- ▶ repeated estimations of each symptomology may lead to double counting of effects as symptoms are often not independent of each other. It is more sensible to estimate overall visual injury and state the symptoms observed. Also no agreement to ERx calculations on visual injury (see answer to #16)

21. para 32 - In OECD guideline 208 it is stated that the seed storage conditions should be described. It should be clarified that a study should not be rejected if the storage conditions are not reported in the cases where the seeds in the control behaved normally in terms of emergence and there is no phytotoxicity measured.



Agreement justification:

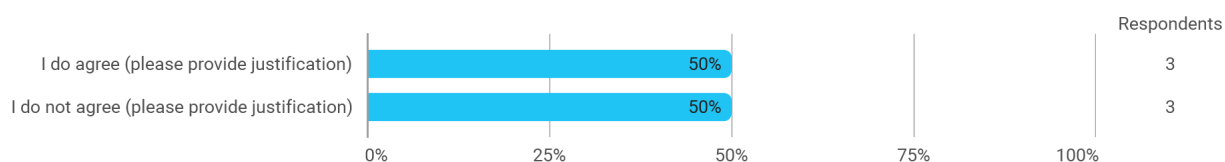
- ▶ internal control
- ▶ validity criteria prevent bad seed batches
- ▶ the health of seed batches is already verified via validity criteria of >70% emergence

- age of seeds and storage conditions are important indicators for quality but most crucial is that the validity criterion of 70% is reached

Disagreement justification:

- no conclusion yet justification: No clarification is needed to perform a risk assessment since the main criterion is plausibility of the results.

22. para 32 -- percent visual injury and qualitative and quantitative description of visual injury (chlorosis, necrosis, wilting, leaf and stem deformation as well as any lack of effects) by the test substance as compared to control plants / a description of the rating scale used to judge visual injury, if visual rating is provided; / stage of the plant development and assessment of visual injury should be supported by pictures; Rational: The use of a phytotoxicity rating system assessing visible herbicide injury in percent affected tissue (e.g. non-green tissue) rather than an injury rating (e.g. rating 1–5, for no, slight, moderate, severe, total plant symptoms, respectively) avoids having to deal with analyzing categorical data. Given a dose-response-relationship ER_x for visual effects compared to un-exposed controls, as required by regulatory authorities, could more easily be derived. Additionally, using the uniform scale of “percent effect” for all endpoints the effects seen are more easily comparable.



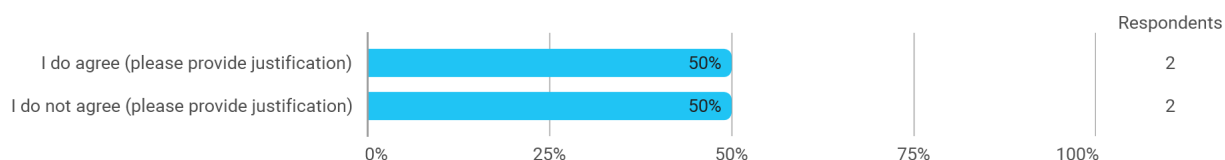
Agreement justification:

- but pictures only at final assessment
- General agreement to ask for this additional information which might help interpreting the results.

Disagreement justification:

- no pictures for GLP
- making pictures of for all plant species would be enormous
- Don't agree to the quantitative aspects, the qualitative aspects are fine. Further agree to necessity of a good description of the rating scale/method. picture should be a "can" not a "must".

23. para 39 - "a description of the rating scale used to judge visual injury that provides the information needed to perform quantitative assessment of the data (e.g. ER_x calculation);"



Agreement justification:

- see point 20: “description of the rating scale used to judge visual injury, that provides the information needed to verify the quantitative assessment of the phytotoxicity data (e.g., ERx calculation)”

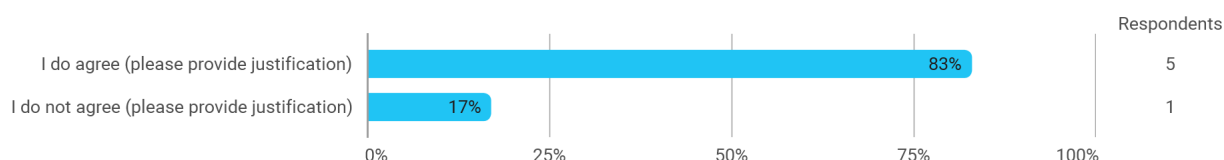
Disagreement justification:

- no ERx calculation for phytotoxicity
- as #22

24. Please complement:

LOER (Lowest Observed Effect Rate) is the lowest rate of the test substance at which effect was observed. In this test, the rate corresponding to the LOER, has a statistically significant effect ($p < 0.05$) within a given exposure period when compared to the control, and is higher than the NOER value.

NOER (No Observed Effect Rate) is the highest rate of the test substance at which no effect was observed. In this test, the rate corresponding to the NOER, has no statistically significant effect ($p < 0.05$) within a given exposure period when compared with the control.



Agreement justification:

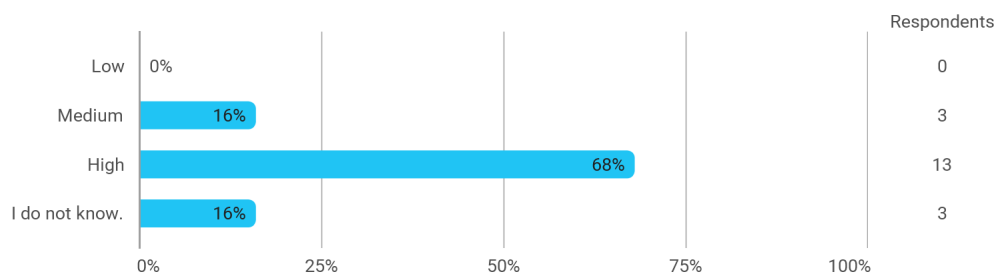
- The amendment is correct since test item units are mostly given as rate, e.g. g/ha.
- the correct dimension in terrestrial plant studies is rate/ha not concentration, hence clarification of endpoints a NOER, LOER, is reasonable

Disagreement justification:

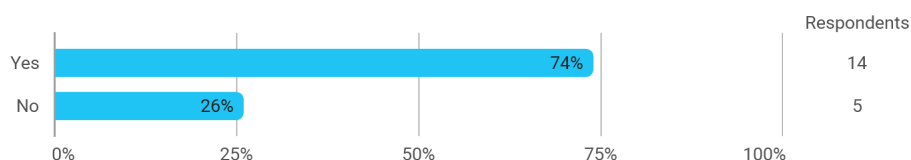
- common sense?

B.12 Test No. 227: Terrestrial Plant Test: Vegetative Vigour Test

Relevance of revision



Interested to be part of an OECD Expert Group

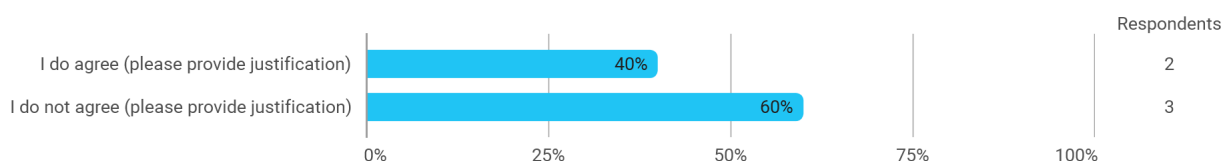


The following suggestions have been collected:

1. New text that should be located under DATA AND REPORTING - statistical analysis:

Results of the control and treatment replicates should be reported “per plant and per pot” so that appropriate statistical analysis can be performed to verify whether the validity criteria regarding the variation in growth is achieved. This new text should be created.

Rational: Results of the control and treatment replicates should be reported “per plant and per pot” so that appropriate statistical analysis can be performed to verify whether the validity criteria regarding the variation in growth is achieved. This new text should be created.



Agreement justification:

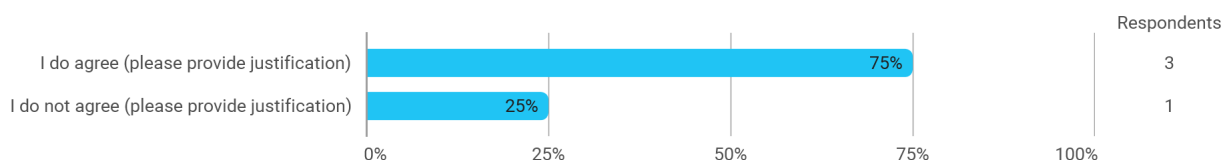
- Additional information which might help reporting the study results in a way adequate for the use in risk assessment.

Disagreement justification:

- NO! per replicate! (per plant is a "fake rep."); due to injury the variation in treated groups can be high - only the variation in control groups should be a validity criteria
- A replicate is defined as a pot and all statistical analyses are performed based on this definition. Individual plants within the same pot are pseudo replicates and not appropriate for the same stats analyses.
- the pot is the replicate, single plants are pseudo replicates.

2. para 2 - Number of seeds per pot. In quite some tests the number of seeds per pot is higher than recommended in the test guideline. Despite this, the plants are performing well and the test can be considered valid. Maybe the recommendation regarding the number of seeds per pot should be adjusted in a certain way.

Rational: Somewhat clearer guidance on this point might lead to less discussions when evaluating the tests.



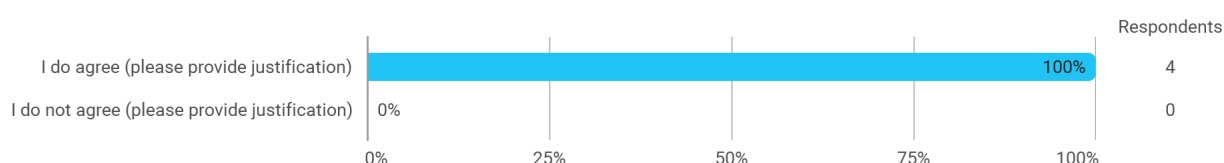
Agreement justification:

- ▶ Number of plants!
- ▶ agree, this part are recommendations to ensure good growth. (External) nutrient supply may be much more crucial for growth condition than space

Disagreement justification:

- ▶ Plant density has an impact on plant growth and thus might affect the results of the study. (refers to points 8-12)

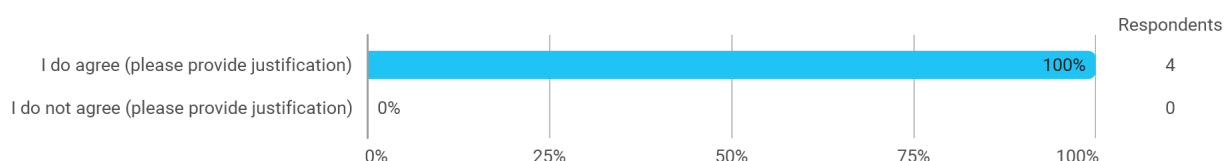
3. **para 3 -“Also, the no observed effect concentration (NOEC), resp. no observed effect rate (NOER), and lowest observed effect concentration (LOEC), resp. lowest observed effect rate (LOER), can be calculated in this test.”**



Agreement justification:

- ▶ Test item units in most studies is “rate”, e.g. g/ha.
- ▶ reflects the rate/ha is the relevant dimension

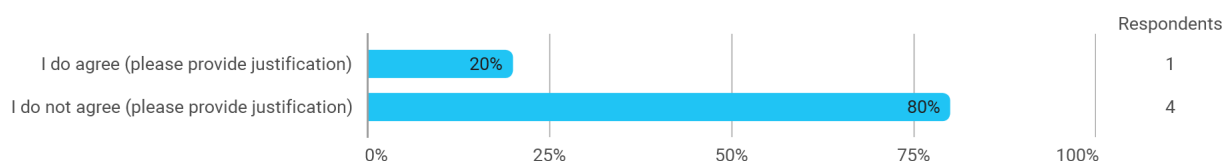
4. **para 3 - Appropriate statistical analysis is used to obtain effective concentration ECx or effective application rate ERx (e.g., EC25, ER25, EC50, ER50) for the most sensitive parameter(s) of interest. Please change “parameter(s)” to “variable(s)”.**
Rational: Variables are measured, parameters are fixed in the experimental design, e.g. light, temperature, test duration, treatments are Parameters.



Agreement justification:

- ▶ Update suggested reflects more recent state of the art approach.

5. **para 5 - The following part should be deleted: - the seedling emergence is at least 70 %;**
Rational: As the emergence is not part of the test endpoints (vegetative vigour study), this validity criterion is not suitable.

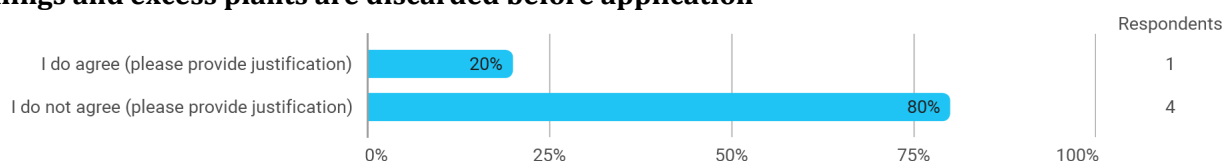


Disagreement justification:

- ▶ it is a quality parameter which is assessed during pre-cultivation
- ▶ this validity criteria ensures that plants were grown from healthy seed batches
- ▶ The criterion ensures that the seeds used in the study are somehow uniform and show a minimum viability. [For crop plants seedling emergence is expected to be > 90%.]
- ▶ (1) good and homogenous emergence is an indicator of good quality and increases likelihood that seedlings are all of good vigour (2) it ensures that seedlings differ only in very few days and show homogenous growth, which is a precondition for good standardization and reduced variability (3) ensures that same batch can be used in both vegetative vigour and seedling emergence studies

6. para 5 - Deletion of the requirement: "the seedling emergence is at least 70 %"

Rational: Validity Criterion for 70% Seedling Emergence is irrelevant as the test itself (i.e. the period after application of a test item) is only started with already emerged seedlings and excess plants are discarded before application

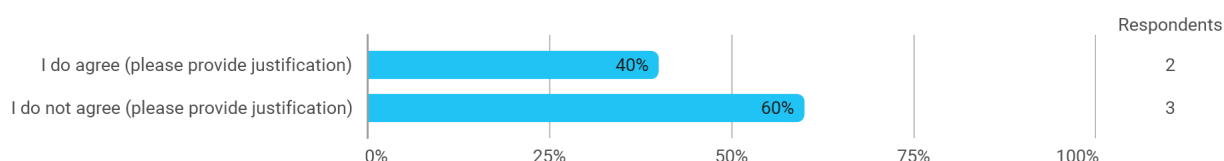


Disagreement justification:

- ▶ see above
- ▶ this validity criteria ensures that plants were grown from healthy seed batches
- ▶ The criterion ensures that the seeds used in the study are somehow uniform and show a minimum viability. [For crop plants seedling emergence is expected to be > 90%.]
- ▶ see #5

7. para 5 - ... the plants do not exhibit visible phytotoxic effects (e.g., chlorosis, necrosis, wilting, leaf and stem deformations) and statistical analysis should support the plants exhibit only normal variation in growth and morphology for that particular species; Please add "and statistical analysis should support" as indicated.

Rational: Since no (statistical) analysis are currently foreseen, this validity criterion is never addressed nor verified in study reports. However, it is of most importance to ensure that seedling have grown in optimal conditions (see also point [10] in the TG 227). Analysis the variation in growth of the control also gives some indications about the statistical power of the test (i.e. limited variation in the controls enhance the probability to identify significant differences in the treatments).



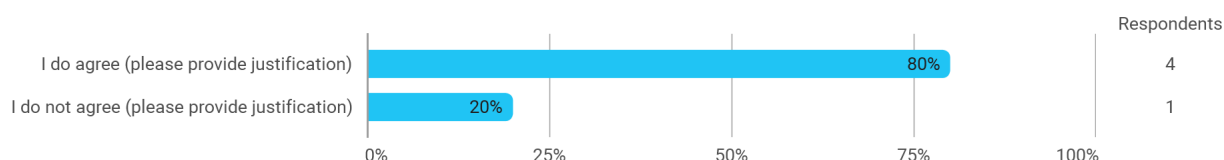
Agreement justification:

- The data on observed visual effects is used in the risk assessment. Preference is given to quantitative data and to ERx derived by statistical analysis - whenever possible (refers to proposals 19-22).

Disagreement justification:

- e.g. a maximum SD or CV for control group would be necessary
- There is not requirement of stats analyses of visual effects
- it concerns qualitative parameters that are assessed by the naked eye. It is not clear what the benefit of the additional effort would be? An experience horticulturist can tell by such assessment whether the plants are in normal conditions or not. Example picture may also provide an impression

8. **para 15 - Number of seeds per pot. In quite some tests the number of seeds per pot is higher than recommended in the test guideline. Despite this, the plants are performing well and the test can be considered valid. Maybe the recommendation regarding the number of seeds per pot should be adjusted in a certain way. Rational: Somewhat clearer guidance on this point might lead to less discussions when evaluating the tests.**



Agreement justification:

- Clarification needed
- agree, this part are recommendations to ensure good growth. (External) nutrient supply may be much more crucial for growth condition than space

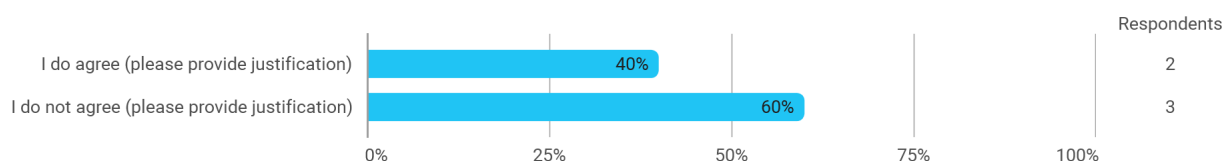
Disagreement justification:

- Plant density has an impact on plant growth and thus might affect the results of the study. (refers to points 2 and 8-12)
9. **para 15 - The number of plants per pot will depend upon the species, pot size and test duration, and should provide adequate and uniform growth conditions and avoid overcrowding and shading of plants by each other for the duration of the test. As an example, one to two corn, soybean, tomato, cucumber, or sugar beet plants per 15 cm container; three rape or pea plants per 15 cm container; and 5 to 10 onion, wheat, or other small seeds per 15 cm container are recommended. Any deviations from these recommendations should be justified. Plant growth should be documented with pictures. Statistical analysis should demonstrate that any deviation does not impact the variability of the measured variables (i.e., increase the variability; see validity criteria). The number of seeds and replicate pots.... Please add the text in bold. Rational: This is the most violated recommendation of the TG.**

Indeed, most of the studies submitted are performed at higher plant density, without any rational about the potential impact of the deviation on the results.

However, there are some indications in the literature that plant density directly impacts plant growth.

Poster Teresiak, SETAC 2015 (30 % weniger Wachstum bei vielen Samen/ kleinem Topf), Poster Simoneit-Gast, SETAC 2015 .



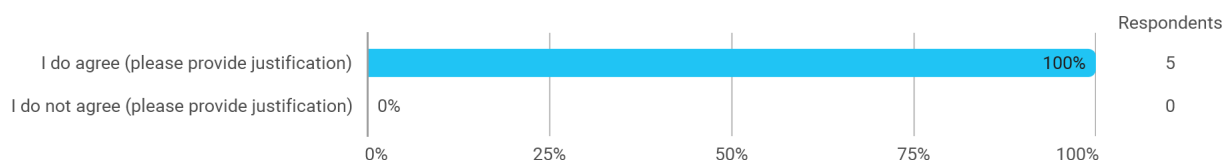
Agreement justification:

- Plant density has an impact on plant growth and thus might affect the results of the study. (refers to points 2 and 8-12)

Disagreement justification:

- see above and no pictures
- this point is a recommendation not a fixed requirements. While for agrochemicals the number of plants/pot have been reduced in recent years, the original recommendations may still be sufficient for studies intended for REACH or Biocide submissions. It would be more useful to clarify conditions for the definition of plant number per pot. This would also serve as guidance when new species are included in studies. Documentation with pictures can be useful, but it should not become a strict requirement (pictures of what exactly, how many, all individuals vs. examples, timings?) It's unclear what parameters of plant growth are meant to be evaluated in statistical analysis.

10. para 15 - "As an example, one to two corn, soybean, tomato, cucumber, or sugar beet plants per 15 cm container; three rape or pea plants per 15 cm container; and 5 to 10 onion, wheat, or other small seeds per 15 cm container are recommended." Please clearly specify the pot size in the examples (i.e. the diameter for a round pot and the side of a square pot). Alternatively, specify the minimum area per seed recommended for each species referred to in the example.

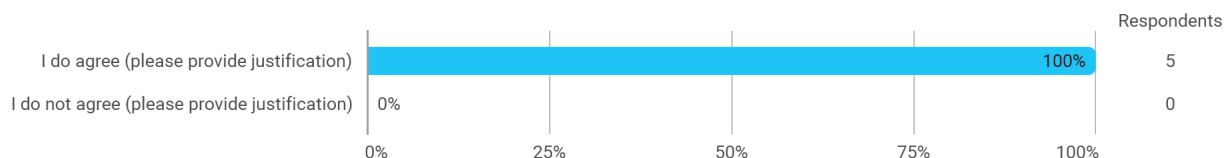


Agreement justification:

- Please clearly specify
- Clarification needed
- Plant density has an impact on plant growth and thus might affect the results of the study. (refers to points 2 and 8-12). Information on pot size is insufficient and inconsistent.
- it is more pragmatic to define plant number/pot by area required/species. Also in line with suggestion #12 nutrient requirements and external nutrient supply are highly important.

sufficient nutrient supply may compensate for reduced area/plant available as growth is more dependent on nutrients than on space (at least during the first 3 weeks of growth)

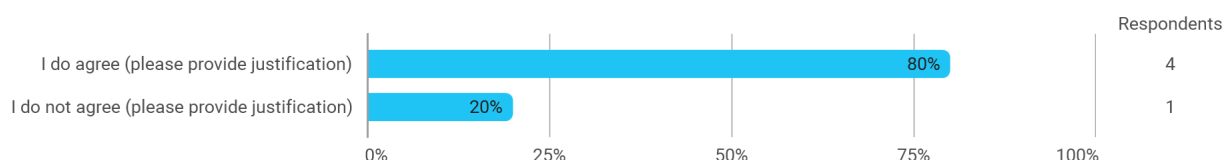
11. para 15 - A defined range for the maximum number of plants per species and pot would be highly appreciated and helpful for a better comparability of the studies.



Agreement justification:

- Clarification needed
- The number of species per pot after thinning is clearly defined as 1. However, information on the maximum number of seeds might be given for more species. The issue “plants per pot” might need to be discussed since plant density has an impact on plant growth and thus might affect the results of the study. (see points 2 and 8-12).
- see #10

12. para 19 - The maximum plant density should be 2 - 5 seeds per container (based on a diameter of 15 cm), based on plant size. As an example, 2 seeds should be used for all dicotyledonous species and maize (Zea maize) and 4 seeds for all monocotyledonous species except of maize, based on a container with a diameter of 15 cm.



Agreement justification:

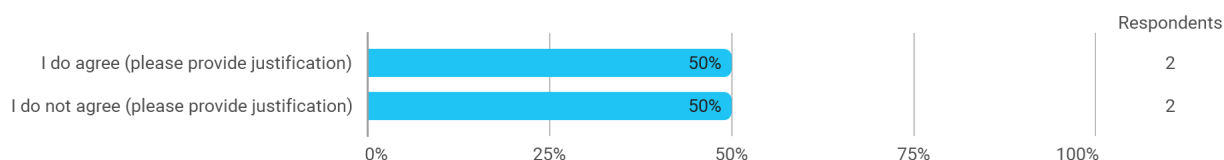
- Please clearly specify
- Clarification needed
- no conclusion yet justification: The proposal refers to para 15. The amendment might not be necessary due to thinning to one plant per pot (paragraph 16). Agree that information on pot size is insufficient and inconsistent. However, the proposal might need to be discussed since plant density has an impact on plant growth and thus might affect the results of the study. (see points 2 and 8-12).

Disagreement justification:

- number of seeds per pot should be adjusted to the size the species can reach within the standard duration of 3 weeks. the nutrient requirements and whether these can be satisfied by external supply also need considerations

13. para 20 - Light intensity: $350 + 50 \mu\text{E}/\text{m}^2\text{s}$.

Rational: The unit given is unclear



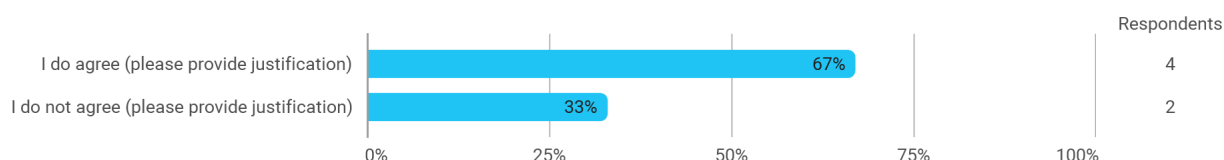
Disagreement justification:

- Light intensity above $400 \mu\text{mol}$ cannot be avoided
- no conclusion yet justification: To conclude on the proposal a justification for the proposal is required.

14. para 20 - Amendment of the range of test conditions which is too narrow.

Rational:

The range of test conditions should not be based on what is technically feasible in a highly controlled growth chamber but on optimal plant growth under greenhouse conditions, and should allow for different requirements of different species. Short deviations from the range should explicitly be allowed, e.g. light intensity is frequently higher than the current range in summer months and the temperature range can be exceeded despite all efforts.



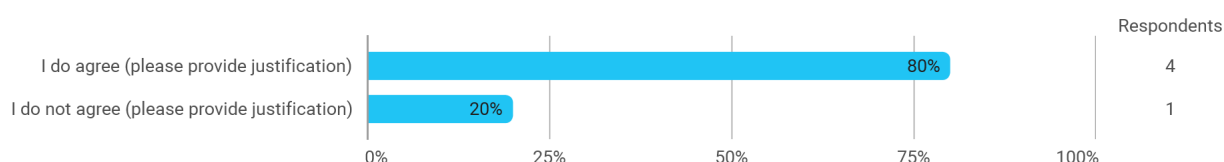
Agreement justification:

- range is indeed too narrow and leads to unnecessary reluctance of acceptance of the study
- but possible with modern technology
- greenhouse conditions are influenced by outside conditions as shading/cooling often can only buffer these with some delay. Hence the range for test conditions needs to be reasonably wide to reflect the technically feasible.

Disagreement justification:

- Glasshouse test are susceptible to local weather conditions. The issue of variation in weather conditions over time and between locations is best addressed via observation of control plant performance and good plant husbandry.
- To conclude on the proposal a justification for the proposal is required.

15. para 22 - Light intensity should be $> 250 \mu\text{mol}/\text{m}^2/\text{s}$; If intensity decreases below $250 \mu\text{mol}/\text{m}^2/\text{s}$, additional lighting is required.



Agreement justification:

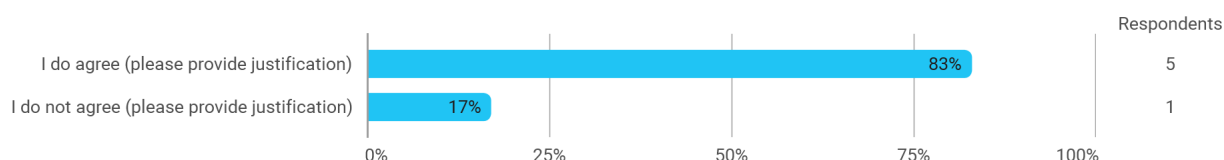
- ▶ Light intensity above 400 μmol cannot be avoided
- ▶ no conclusion yet justification: To conclude on the proposal a justification for the proposal is required. Please note that the proposal actually refers to para 20.
- ▶ Okay

Disagreement justification:

- ▶ currently min light intensity is 300 $\mu\text{mol}/\text{m}^2/\text{s}$

16. para 25 - Method for measurement and interpretation of visual injury has to be better and more precisely described / It needs to be clarified that symptoms and intensity of visual injury compared to the control should be presented in the report, but no statistical analysis is to be performed.

Rational: Visual Injury is a very subjective measure and not fit for statistical analysis. So far, studies from different laboratories have performed the Visual Assessment in different ways (e.g. scale ratings, different symptoms, separate assessment of symptoms or combined etc.); none of the methods allows for statistical analysis and objective measures like dry weight or survival are by far more adequate for effect value derivation



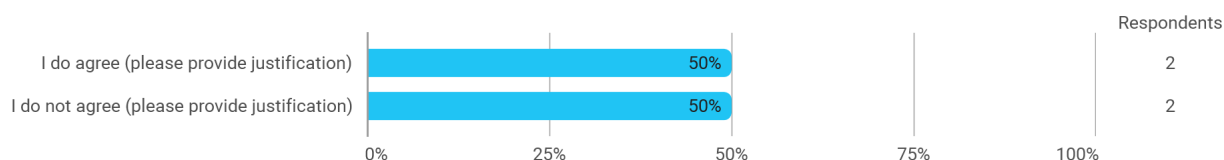
Agreement justification:

- ▶ stat. analysis of Visual Injury is not reliable
- ▶ Please clearly specify
- ▶ Clarification needed
- ▶ valuable clarification. method can be better described. visual injury is a qualitative endpoint besides a set of quantitative endpoints in this guideline. it should be clear that statistic is calculated for quantitative parameter but not for qualitative ones. this is in line with other international guidance (e.g. OCPPS Guidance 850.4000)

Disagreement justification:

- ▶ The data on observed visual effects is used in the risk assessment. Preference is given to quantitative data and to ERx derived by statistical analysis - whenever possible (refers to proposals 19-22).

17. para 26 - “If the NOEC or NOER is to be estimated, more replicates may be needed to obtain the desired statistical power (23).”



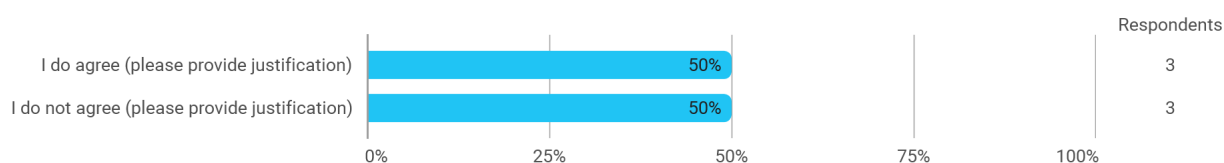
Agreement justification:

- The unit of the test item in most studies is “rate”, e.g. g/ha.

Disagreement justification:

- A compromise between the number of replicates and the number of treatments must be found.
- main scope of studies is derivation of an ER50/ER25

18. para 27 - It is suggested to introduce reproductive effects endpoints (e.g. flowering, seed production), as these sensitive effects on plants are not yet included in the risk assessment for non-target terrestrial plants. This will subsequently require updating of data requirements / guidance documents to enable a full plant risk assessment, covering the full lifecycle of plants. see Strandberg et al, 2019: Strandberg, B. et al, 2019. Pesticide effects on non-target terrestrial plants at individual, population and ecosystem level (PENTA), Danish Environmental Protection Agency, Pesticide Research 182, September 2019. available at <https://www2.mst.dk/Udgiv/publications/2019/09/978-87-7038-111-6.pdf>



Agreement justification:

- independent guideline necessary
- Extension of the studies on juveniles to gain reproductive effects endpoints is feasible – in addition or as supplement to the existing guidelines.

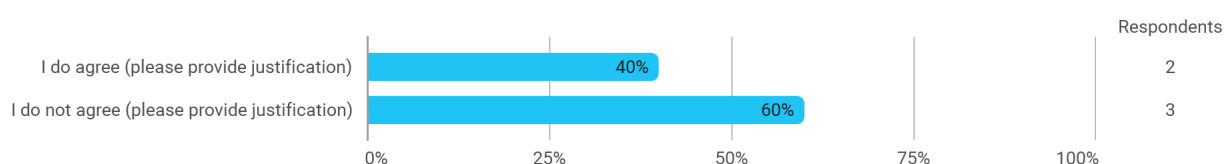
Disagreement justification:

- reproduction testing is not part of the veg vig test and should be addressed separately if required
- The issue of measuring effects on plant reproductive parameters is currently being evaluated by a WG working under the auspices of the SETAC Plant Interest Group. In the event that a TG is required for this purpose, methods currently described in OECD 227 will not be adequate, due to differences in test duration, need for increased nutrient supply and the fact that many species currently covered by TG 227 are not readily amenable to reproductive testing eg biennials such as sugar beet or fruiting veg such as tomatoes.

- This protocol is designed for early growth of seedlings. A protocol fit for purpose for the assessments of reproductive endpoints will require major adaptations. Further such protocols are expected to be species specific as requirements of plants deviate much more in mature stages than during early growth. If reproductive endpoints should be assessed in OECD non-target plant studies this should be done under independent test guideline yet to be developed. Regarding potential relevance of reproductive endpoints I would like to point to the review by Christl et al. 2019 (DOI: 10.1002/ieam.4218; Comparative Assessment of Vegetative and Reproductive Terrestrial Plant Species Endpoints from Exposure to Herbicides and Potential Environmental Implications: A Review)

19. para 27 – A better quantification of phytotoxic effects is recommended. It should be evaluated by use of digital image quantification.

Rational: At the moment phytotoxicity is quantified by visual inspection, which can be affected by operator's subjectivity.



Agreement justification:

- The observed visual effects are to be quantified - where possible - and an ERx derived. (refers to proposals 19-22).

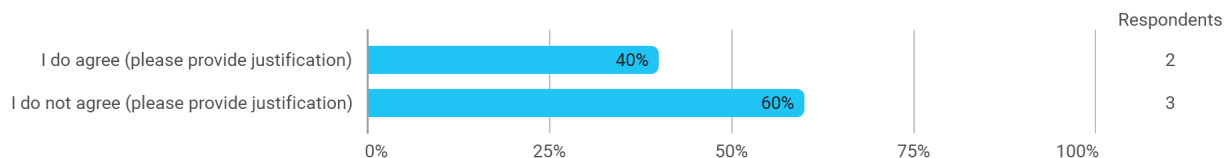
Disagreement justification:

- the plant damage is very individually between the different test species depending the different test items
- Visual effects invariably translate into effects on biomass and shoot length. These parameters are already assessed and provide a qualitative, objective assessment of effects. In contrast, visual scores are subjective, lack granularity and do not provide a measure of variability. While digital imaging may provide a solution, it is time consuming expensive and highly uncertain whether estimation of endpoints based on visual observations will add value to risk assessment given that this TG already generates several conservative endpoints.
- This suggestion seems quite technocratic. Such sophisticated equipment is very expensive and I am not aware of any contract laboratory running OECD 208 and 227 studies that does have such an equipment in their greenhouses. Besides, I am not quite clear what the added value of such information would be for the outcome of the study and a risk assessment?

20. para 27 - "A uniform scoring system for visual injury should be used to evaluate the observable toxic responses. Examples for performing qualitative and quantitative visual ratings are provided in references (23) (24).As indicated there, preferably qualitative data (e.g. from 0 to 100 %) should be used to allow ERx calculation when multiple rate testing has been conducted to generate a dose-response curve."

Rational: The use of a phytotoxicity rating system assessing visible herbicide injury in percent affected tissue (e.g. non-green tissue) rather than an injury rating (e.g. rating 1–5, for no, slight, moderate, severe, total plant symptoms, respectively) avoids having to deal with analyzing categorical data. Given a dose-response-relationship ERx for

visual effects compared to un-exposed controls, as required by regulatory authorities, could more easily be derived. Additionally, using the uniform scale of “percent effect” for all endpoints the effects seen are more easily comparable.



Agreement justification:

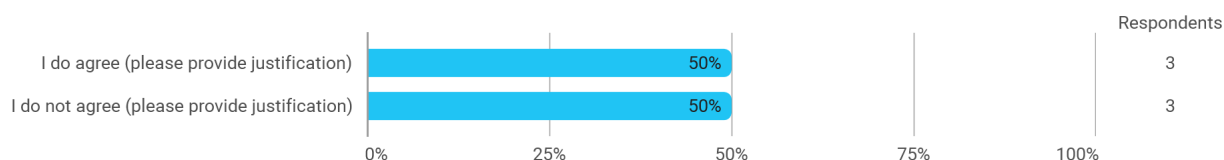
- The observed visual effects are to be quantified - where possible - and an ERx derived. (refers to proposals 19-22).

Disagreement justification:

- No ERx calculation for phytotoxicity
- Disagree: It is uncertain whether estimation of endpoints based on visual observations will add value to risk assessment given that this TG already generates several endpoints. The issue of visual score endpoints is currently being evaluated by a WG working under SETAC Plant Interest Group.
- visual injury is a qualitative parameter mainly needed for the validity criterion of "no phytotoxicity in the control" and to ensure that late onset of injury is detected (->prolongation). Visual estimations of injury are variable between individual assessors and along the scoring scale (larger in the middle range). The Guideline already includes several quantitatively measured parameter, hence there is no need to squeeze a qualitative parameter artificially into quasi-quantitative frame.

21. para 27 - For phytotoxicity, both NTTP guidelines (OECD 208, 227) need to explain how to report phytotoxicity clearly so that an endpoint can be derived. In cases where the phytotox endpoint is lower than the regular study endpoint, the phytotox endpoint should be considered for risk assessment.

Rational: Phytotoxicity is an important parameter to take into account in the derivation of the endpoint from the test. Therefore it is needed that the phytotoxicity is reported in a clear way, so that an endpoint can be derived based on this parameter.



Agreement justification:

- Please clearly specify
- The observed visual effects are to be quantified - where possible - and an ERx derived. (refers to proposals 19-22).

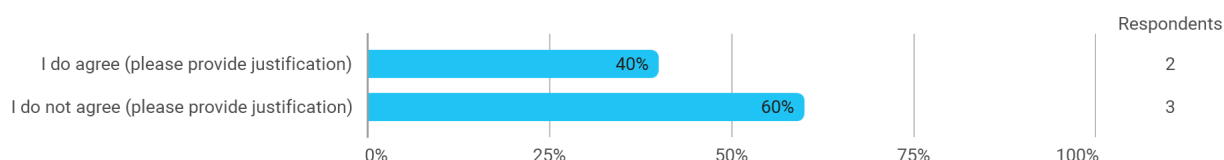
Disagreement justification:

- Unreliable and subjective endpoint with unknown ecological relevance should not be used instead of a reliably measurable endpoint like biomass

- Disagree: It is uncertain whether estimation of endpoints based on visual observations will add value to risk assessment given that this TG already generates several endpoints. The issue of visual score endpoints is currently being evaluated by a WG working under SETAC Plant Interest Group.
- is it really suitable to use a naked eye estimation with high variability to override a meaningful and robustly measured endpoint

22. para 27 - A uniform scoring system for visual injury should be used to evaluate the observable toxic responses. Examples for performing qualitative and quantitative visual ratings are provided in references (12) (25). As indicated there, preferably quantitative data (e.g. from 0 to 100 %) should be used to allow ERx calculation when multiple rate testing has been conducted to generate a dose-response curve.

Rational: The use of a phytotoxicity rating system assessing visible herbicide injury in percent affected tissue (e.g. non-green tissue) rather than an injury rating (e.g. rating 1–5, for no, slight, moderate, severe, total plant symptoms, respectively) avoids having to deal with analyzing categorical data. Given a dose-response-relationship ERx for visual effects compared to un-exposed controls, as required by regulatory authorities, e.g. for the authorisation of pesticides in Europe, could more easily be derived. Additionally, using the uniform scale of “percent effect” for all endpoints the effects seen are more easily comparable.



Agreement justification:

- The observed visual effects are to be quantified - where possible - and an ERx derived. (refers to proposals 19-22).

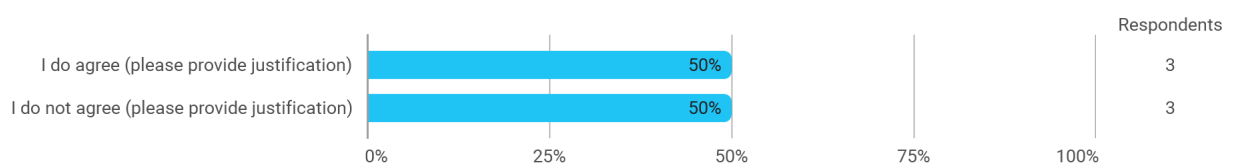
Disagreement justification:

- No ERx calculation for phytotoxicity
- Disagree: It is uncertain whether estimation of endpoints based on visual observations will add value to risk assessment given that this TG already generates several endpoints. The issue of visual score endpoints is currently being evaluated by a WG working under SETAC Plant Interest Group.
- visual injury is a qualitative parameter mainly needed for the validity criterion of "no phytotoxicity in the control" and to ensure that late onset of injury is detected (->prolongation). Visual estimations of injury are variable between individual assessors and along the scoring scale (larger in the middle range). The Guideline already includes several quantitatively measured parameter, hence there is no need to squeeze a qualitative parameter artificially into quasi-quantitative frame.

23. para 29 - "For the growth of the seedlings (weight and height) as continuous endpoints ECx or ERx and its confidence limits can be estimated by using appropriate regression analysis (e.g., Bruce-Versteeg non-linear regression analysis (26)). Note

that the biomass of individuals that died during the study should be set to zero and included in the dose-response analysis."

Rational: Dose-response models used to derive ERx values for biomass measurements (i.e. shoot dry and fresh weight as well as shoot height) typically encompass effect sizes from 0% to 100%. In terms of an 100% effect in a treatment compared to control, this means that there is zero biomass in that treatment, which can only be achieved if the plants are fully mineralized. As this is impossible due to the study design, the closest approximation to an effect size of 100% is given by setting the plants to zero that died during the experiment. However, this is not explicitly mentioned in the OECD GD 227, which results in statistical evaluation of biomass effects in which dead individuals are excluded from dose-response analyzes. This, in turn, violates statistical rules and subsequently leads to false endpoints for environmental risk assessment, which constitute a serious problem and therefore should be corrected as soon as possible by including the suggested sentence above.



Agreement justification:

- ▶ as well for plant height
- ▶ Plants cannot be defined as dead (roots are not assessed). Therefore, assessment parameters should only be set to zero when there is an absence of measurable biomass.

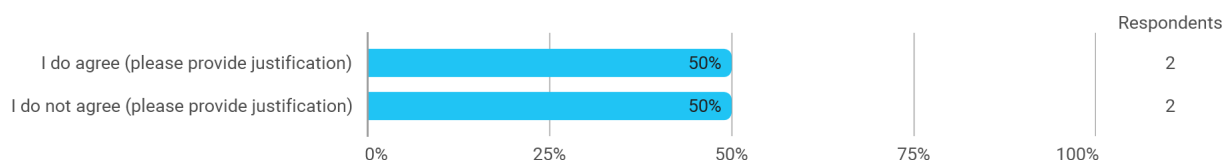
Disagreement justification:

- ▶ no inclusion of dead plants in endpoint derivation other than survival
- ▶ Disagree to the sentence "Note that the biomass of individuals that died during the study should be set to zero and included in the dose-response analysis"(1) Including dead individuals in evaluation of weight seems to be questionable at all, since evaluation of mortality should not be mixed up with evaluation of growth. (2) How to define "dead" in the context of plants? (3) A metric variable such as weight or height can never be zero. If there is anything measurable, then there is a number available. If there is nothing measurable, then it is a missing value – not zero. Zero is an arbitrary assumption, not a measurement result. Setting metric data to zero will result in a significant reduction of the treatment mean. Thereby, mortality is mixed up with growth, hence, results of different studies are no longer comparable.
- ▶ no conclusion yet - General agreement that excluding the most severely affected individuals from ERx determination biases the results. However, following that proposal different effect parameters are mixed up. A discussion on this item is appreciated.

24. para 29 - For the growth of the seedlings (weight and height) as continuous endpoints ECx or ERx and its confidence limits can be estimated by using appropriate regression analysis (e.g., Bruce-Versteeg non-linear regression analysis (25)) by using three-parametric models, which include the control as additional parameter.

Rational: It is common practice in the analysis of continuous biomass data (i.e. shoot dry and fresh weight as well as shoot height) that they are rescaled by calculating the

effect size in percent. Rescaling of data is considered as convenient because two-parametric dose-response models can be used for dose-response analysis. However, this is not state of the art anymore and is also not in line with the OCED GD 54 “Current Approaches in the Statistical Analysis of Ecotoxicity Data”, since rescaling reduce the variation of the data by excluding the natural variability (i.e. control values are missing in dose-response analysis). Moreover, two-parametric models generally have a poorer fit to continuous data compared to three-parametric models, which affects the precision and accuracy of derived ERx values. Therefore, any ad hoc adjustments of continuous data before fitting a dose-response model is inappropriate and should be avoided.



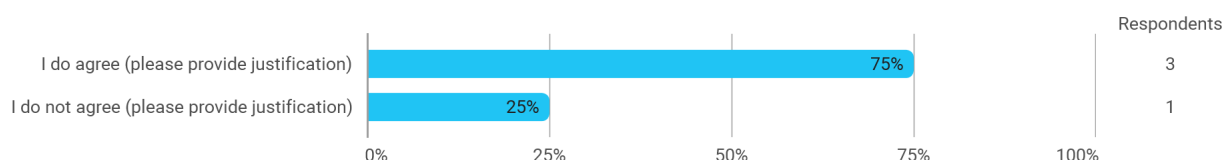
Agreement justification:

- Update suggested reflects state of the art.

Disagreement justification:

- The optimal model should always be weighed up individually and cannot be generalised. The calculated end points must match the visual impression in the greenhouse!

25. para 29 - Temperature for determination of constant weight of plant shoots should be given as a range, i.e. 60 ± 5 °C



Agreement justification:

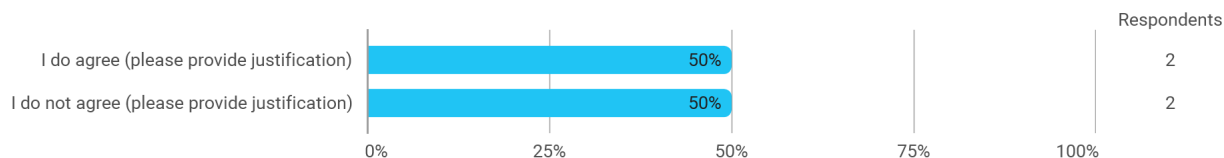
- but how is constant weight defined? e.g. deviation <1% after 3 hours

Disagreement justification:

- no conclusion yet justification: A range for drying temperature might be helpful but the range proposed is too high to allow standardisation.

26. para 29 - “For the growth of the seedlings (weight and height) as continuous endpoints ECx or ERx and its confidence limits can be estimated by using appropriate regression analysis (e.g., Bruce-Versteeg non-linear regression analysis (26)) by using three-parametric models, which include the control as additional parameter.” Rational: Dose-response models used to derive ERx values for biomass measurements (i.e. shoot dry and fresh weight as well as shoot height) typically encompass effect sizes from 0% to 100%. In terms of an 100% effect in a treatment compared to control, this means that there is zero biomass in that treatment, which can only be achieved if the plants are fully mineralized. As this is impossible due to the study design, the closest approximation to an effect size of 100% is given by setting the plants to zero that died during the experiment. However, this is not explicitly

mentioned in the OECD GD 227, which results in statistical evaluation of biomass effects in which dead individuals are excluded from dose-response analyzes. This, in turn, violates statistical rules and subsequently leads to false endpoints for environmental risk assessment, which constitute a serious problem and therefore should be corrected as soon as possible by including the suggested sentence above.



Agreement justification:

- Update suggested reflects state of the art.

Disagreement justification:

- The optimal model should always be weighed up individually and cannot be generalised. The calculated end points must match the visual impression in the greenhouse

27. para 30 - percent visual injury and qualitative and quantitative description of visual injury. Replace by: - a qualitative description and a quantitative assessment of the visual injuries (i.e., ERx calculation). - description of the rating scale used to judge visual injury that provides the information needed to verify the quantitative assessment of the phytotoxicity data (e.g., ERx calculation);- ECx or ERx (e.g., EC50, ER50, EC25, ER25) values and related normalised confidence limits.

Rational: - As mentioned under 2., an endpoint can be calculated for different biological variables. Therefore, it is important that ERx values are reported for each variable measured in the test. This includes visual injuries.

This is supported by the fact that visual injury (i.e., phytotoxicity) is clearly mentioned in the TG as relevant variable:

i) in the principle of the test as being a relevant endpoint: “Endpoints measured are visual assessment of seedling emergence, dry shoot weight (alternatively fresh shoot weight) and in certain cases shoot height, as well as an assessment of visible detrimental effects on different parts of the plant” (p. 1).

ii) in the ECx definition in Annex 1:

ECx. x% Effect Concentration or ERx. x% Effect Rate is the concentration or the rate that results in an undesirable change or alteration of x% in the test endpoint being measured relative to the control (e.g., 25% or 50% reduction in seedling emergence, shoot weight, final number of plants present, or increase in visual injury would constitute an EC25/ER25 or EC50/ER50 respectively)."

Although plant visual injury is commonly measured and reported in almost all studies, this variable is not always assessed in a quantitative way, i.e. ERx values derived. In study reports the assessment of this variable is often limited to a quantitative description of the effects observed.

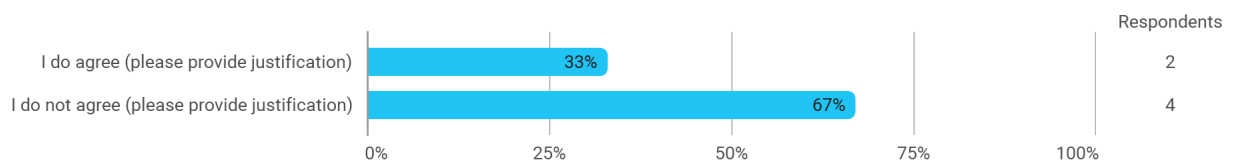
However, for the risk assessment of plant protection products an agreement has recently been met at EU level and between all Member States of the European central Zone to consider this effect in a quantitative way in the risk assessment (see references below). Therefore, it is of most importance that ERx values are now calculated.

The report of the normalized confidence limits (i.e., confidence limits / the corresponding ERx values) is more informative.

Publications:

EFSA Supporting publication 2019:EN-1673

Central Zone Evaluation Manual, a Working document on Risk Assessment of Plant Protection Products in the Central Zone - Ecotoxicology. Version 1.0, May 2021.



Agreement justification:

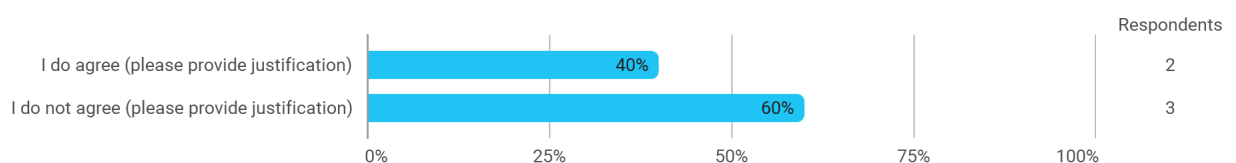
- The information is necessary to evaluate the test.

Disagreement justification:

- stat. analysis of Visual Injury is not reliable
- No ERx calculation for phytotoxicity
- It is uncertain whether estimation of endpoints based on visual observations will add value to risk assessment given that this TG already generates several endpoints. The issue of visual score endpoints is currently being evaluated by a WG working under SETAC Plant Interest Group.
- visual injury is a qualitative parameter mainly needed for the validity criterion of "no phytotoxicity in the control" and to ensure that late onset of injury is detected (->prolongation). Visual estimations of injury are variable between individual assessors and along the scoring scale (larger in the middle range). The Guideline already includes several quantitatively measured parameter, hence there is no need to squeeze a qualitative parameter artificially into quasi-quantitative frame.

28. para 30 - a description of the rating scale used to judge visual injury that provides the information needed to perform quantitative assessment of the data (e.g., ERx calculation)".

Rational: The use of a phytotoxicity rating system assessing visible herbicide injury in percent affected tissue (e.g. non-green tissue) rather than an injury rating (e.g. rating 1–5, for no, slight, moderate, severe, total plant symptoms, respectively) avoids having to deal with analyzing categorical data. Given a dose-response-relationship ERx for visual effects compared to un-exposed controls, as required by regulatory authorities, could more easily be derived. Additionally, using the uniform scale of “percent effect” for all endpoints the effects seen are more easily comparable.



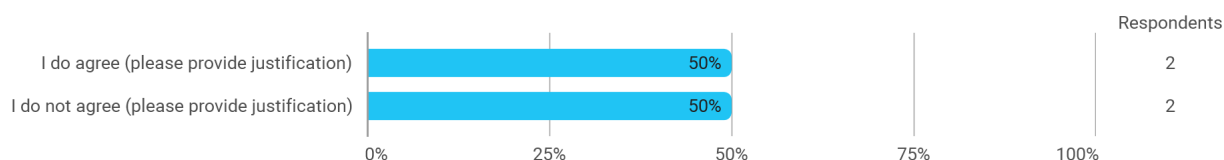
Agreement justification:

- The information is necessary to evaluate the test.

Disagreement justification:

- ▶ No ERx calculation for phytotoxicity
- ▶ It is uncertain whether estimation of endpoints based on visual observations will add value to risk assessment given that this TG already generates several endpoints. The issue of visual score endpoints is currently being evaluated by a WG working under SETAC Plant Interest Group.
- ▶ visual injury is a qualitative parameter mainly needed for the validity criterion of "no phytotoxicity in the control" and to ensure that late onset of injury is detected (->prolongation). Visual estimations of injury are variable between individual assessors and along the scoring scale (larger in the middle range). The Guideline already includes several quantitatively measured parameter, hence there is no need to squeeze a qualitative parameter artificially into quasi-quantitative frame.

29. para 30 - percent visual injury and qualitative and quantitative description of visual injury (chlorosis, necrosis, wilting, leaf and stem deformation as well as any lack of effects) by the test substance as compared to control plants; - a description of the rating scale used to judge visual injury, if visual rating is provided;" - stage of the plant development and assessment of visual injury should be supported by pictures.



Agreement justification:

- ▶ The information is necessary to evaluate the test.

Disagreement justification:

- ▶ no pictures
- ▶ visual injury is a qualitative parameter mainly needed for the validity criterion of "no phytotoxicity in the control" and to ensure that late onset of injury is detected (->prolongation). Visual estimations of injury are variable between individual assessors and along the scoring scale (larger in the middle range). The Guideline already includes several quantitatively measured parameter, hence there is no need to squeeze a qualitative parameter artificially into quasi-quantitative frame.

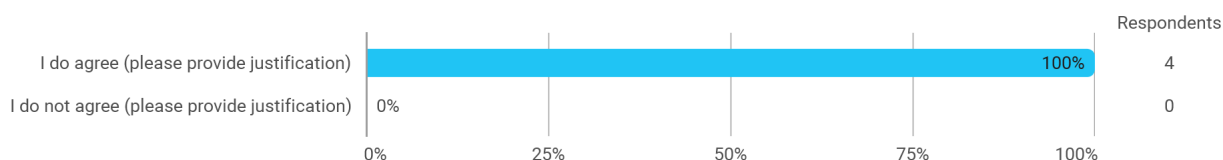
30. Annex 1 definition: Please complement: LOER (Lowest Observed Effect Rate) is the lowest rate of the test substance at which effect was observed. In this test, the rate corresponding to the LOER, has a statistically significant effect ($p < 0.05$) within a given exposure period when compared to the control, and is higher than the NOER value. NOER (No Observed Effect Rate) is the highest rate of the test substance at which no effect was observed. In this test, the rate corresponding to the NOER, has no statistically significant effect ($p < 0.05$) within a given exposure period when compared with the control.

Phytotoxicity: please add to the definition that phytotoxicity includes all endpoints assessed according to this Test GL and must not be confused by "visual injury" or

"visible detrimental symptoms" which are often simply referred to as (visual) "phytotoxicity".

Visual assessment: please add to the definition that the outcome of this assessment should preferably be referred to as "visual injury" instead of "phytotoxicity" because the visually assessed symptoms are only one part of the phytotoxicity symptomology.

Rational: there is much confusion in about "phytotoxicity" whether being referred to as simply "visual injury" or overall phytotoxicity including also the measured endpoints like biomass, height and survival.



Agreement justification:

- ▶ Please clearly specify
- ▶ The proposed definitions is helpful explanatory information.
- ▶ reflects that the relevant dimension for NTTP testing is rate/ha. Clarifies the distinction between phytotoxicity and visual injury

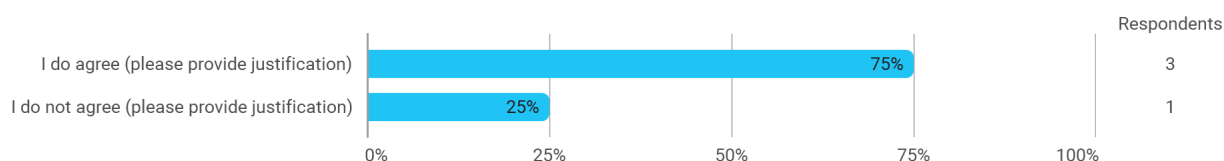
31. Annex 3: The list of potential test species should be adapted with clear recommendation for identification of suitable seed batches of non-standard species. Further recommendations for the assessment of growth characteristics and homogeneity of growth would be helpful to identify whether sufficient standardization is possible with this species. Alternatively or in addition, a statement could be included that acknowledges the lack of ring-testing of non-crop species suitability.

Rational: The suggested change is linked to §10.

Standardization in plant testing is getting under scrutiny by claims that wild species would be more sensitive and more relevant for testing. However, standardization is an important principle of toxicity tests to make results comparable between different substances. For lower tiers (where this TGL applies) test species are chosen based on their availability (seed sources) and suitability as surrogate species under standardized test conditions. High variability in germination / emergence as often seen in wild species hampers standardization. This should be acknowledged in the Test GL. Further, regarding the relevance of test species, crop species may be target and non-target species for herbicides: non-targets on adjacent fields, target on the same field if occurring as unwanted volunteer crops among a different crop species. It is suggested to include references to Pallett et al. 2007 (study on suitability of non-crop species) and Christel et al. 2019 (review on sensitivity of crop vs. non-crop species).

Pallett et al. (2007). Performance of potential non-crop or wild species under OECD 208 testing guideline study conditions for terrestrial non-target plants. PestManagSci, 63:134-140

Christel et al. (2019). Comparative Assessment of the Intrinsic Sensitivity of Crop Species and Wild Plant Species to Plant Protection Products and Their Active Substances and Potential Implications for the Risk Assessment: A Literature Review. IEAM, 15(2):176-189.



Agreement justification:

- yes, but too much restriction can also make things more difficult
- additional guidance on quality parameters for the choice of additional species would be appreciated

Disagreement justification:

- no conclusion yet justification: feasibility and necessity are doubted

32. Annex 4, §2, 3:

2) The test assesses the potential effects on plants following deposition of the test substance on the leaves and above-ground portions of plants. Plants are grown from seed usually to the 2- to 4- true leaf stage. Test substance is then sprayed on the plant and leaf surfaces at appropriate rate(s). After the application, the plants are evaluated against untreated control plants for effects on vigour and growth at various time intervals through 21 - 28 days from treatment. A test period of 21 days can be sufficient for the 10 crop species listed in Annex 4. Endpoints measured are dry shoot weight (alternatively fresh shoot weight), and in certain cases shoot height, as well as a qualitative or semi-quantitative assessment of visible detrimental effects on different parts of the plants. These measurements and observations are compared to those of untreated control plants.

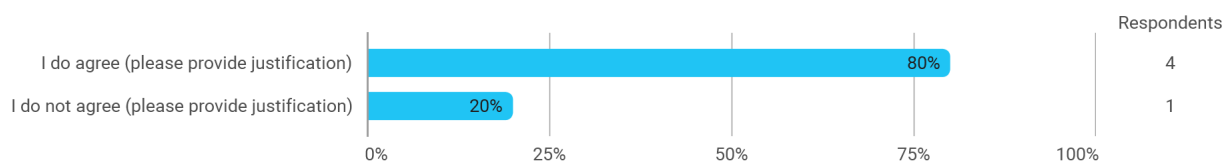
3) The test can be conducted in order to determine the dose-response curve, or at a single concentration/rate as a limit test according to the aim of the study. If results from the single concentration/rate test exceed a certain toxicity level (e.g. whether effects greater than x % are observed), a range finding test is carried out to determine upper and lower limits for toxicity followed by a multiple concentration/rate test to generate a dose-response curve. For quantitatively measured endpoints, appropriate statistical analysis are used to obtain an effective concentration EC_x or an effective application rate ER_x (e.g. EC₂₅, ER₂₅, EC₅₀, ER₅₀) for the most sensitive parameter(s) of interest. Also, the no observed effect concentration (NOEC) and lowest observed effect concentration (LOEC) can be calculated in this test. Some of the data generated in phytotoxicity tests are qualitative, such as ratings based upon visual symptoms of phytotoxicity. Qualitative data such as ratings are not statistically analysed for these tests but may be used to report qualitative no-effect levels.

Rational:

2) The test assesses the potential effects on plants following deposition of the test substance on the leaves and above-ground portions of plants. Plants are grown from seed usually to the 2- to 4- true leaf stage. Test substance is then sprayed on the plant and leaf surfaces at appropriate rate(s). After the application, the plants are evaluated against untreated control plants for effects on vigour and growth at various time intervals through 21 - 28 days from treatment. A test period of 21 days can be sufficient for the 10

crop species listed in Annex 4. Endpoints measured are dry shoot weight (alternatively fresh shoot weight), and in certain cases shoot height, as well as a qualitative or semi-quantitative assessment of visible detrimental effects on different parts of the plants. These measurements and observations are compared to those of untreated control plants.

3) The test can be conducted in order to determine the dose-response curve, or at a single concentration/rate as a limit test according to the aim of the study. If results from the single concentration/rate test exceed a certain toxicity level (e.g. whether effects greater than x % are observed), a range finding test is carried out to determine upper and lower limits for toxicity followed by a multiple concentration/rate test to generate a dose-response curve. For quantitatively measured endpoints, appropriate statistical analysis are used to obtain an effective concentration ECx or an effective application rate ERx (e.g. EC25, ER25, EC50, ER50) for the most sensitive parameter(s) of interest. Also, the no observed effect concentration (NOEC) and lowest observed effect concentration (LOEC) can be calculated in this test. Some of the data generated in phytotoxicity tests are qualitative, such as ratings based upon visual symptoms of phytotoxicity. Qualitative data such as ratings are not statistically analyzed for these tests but may be used to report qualitative no-effect levels.



Agreement justification:

- ▶ stat. analysis of Visual Injury is not reliable
- ▶ addition to paragraph 3 is in line with other international Guidance, e.g. OCPPS Guidance 850.4000 (Background and Special Considerations – Tests with Terrestrial and Aquatic plants, Cyanobacteria, and Terrestrial soil Core Microcosms)

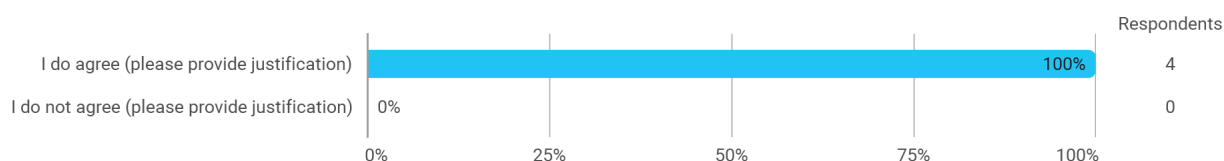
Disagreement justification:

- ▶ ANNEX 4 gives an example for appropriate growth conditions in growth chambers for certain crop species listed therein. In contrast to that, the proposed additional §2, 3 provides some kind of summary of the test design, which is not necessary. Also, visual injury should be quantified.

33. Annex 4: modify reference to Annex 4: "The test conditions should approximate those conditions necessary for normal growth for the species and varieties tested (Annex 4 provides examples of test condition)." To "The test conditions [...] varieties tested (Annex 4 provides an examples of test conditions in growth chambers compared to greenhouses outlined below)."

Rational: modify reference to Annex 4:

"The test conditions should approximate those conditions necessary for normal growth for the species and varieties tested (Annex 4 provides examples of test condition)." to "The test conditions [...] varieties tested (Annex 4 provides an examples of test conditions in growth chambers compared to greenhouses outlined below)."

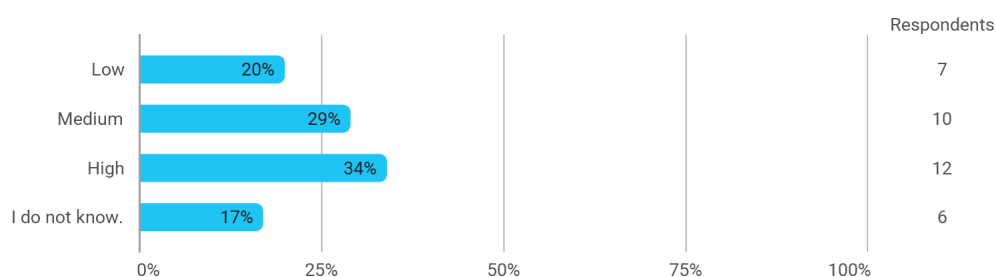


Agreement justification:

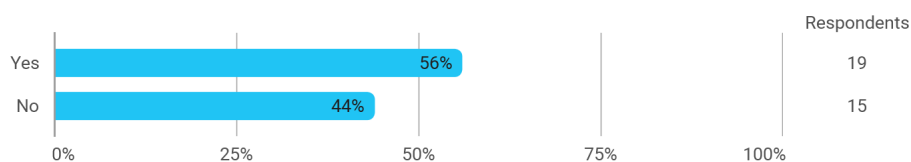
- The modification proposed actually refers to paragraph 19.
- It's relevant. We had lengthy discussions in the past with an authority that wanted to invalidate plant studies conducted in Greenhouses because the narrow conditions for growth chambers (Annex 4) were not complied with. However, the correction could be simpler, e.g. para 19 in OECD 227 "as an example [for condition recommendations in growth chambers] see Annex 4. And in Annex 4 a simple statement that these narrow conditions are suitable for test in greenhouses.

B.13 Test No. 207: Earthworm, Acute Toxicity Tests

Relevance of revision

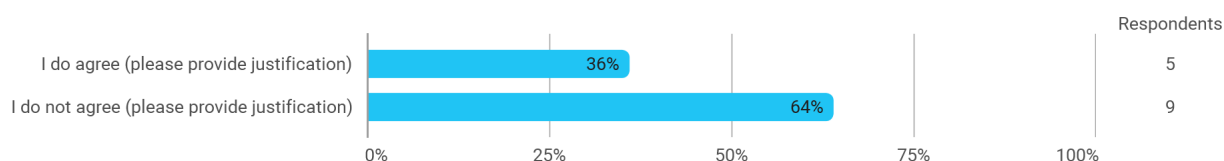


Interested to be part of an OECD Expert Group



The following suggestions have been collected:

1. **General:** It might be considered to make the requirement for analytical check for the behaviour of the chemical compulsory. The comment might be relevant for all tests using artificial soil.
Rational: It is mandatory for aquatic organisms (even for acute tests), including sediment dwellers. Indeed, the dissipation of the chemical largely influences the effect, therefore the endpoint get from the test (which is the main aim). The dissipation rate can considerable different in the test media than in the field. Therefore, the use of the endpoints to compare the toxicity of the different chemicals might be a questionable practice. What the endpoint actually express is a combination of the toxicity and the dissipation rate of the chemical in the test media.



Agreement justification:

- for application verification

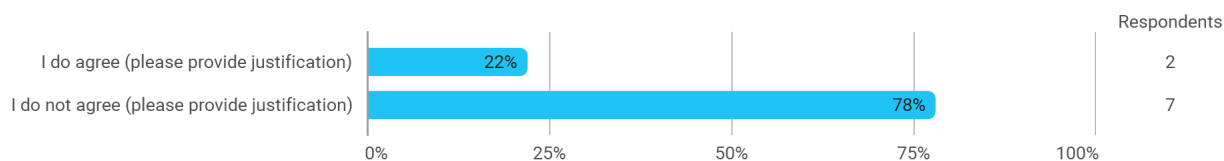
Disagreement justification:

- A dose verification should not be obligatory
- Analytical measurements in GLP soil tests are not meaningful due to the analytical difficulties in the soil matrix. Only radioactive measurements would provide reasonable results because of multiple interactions between pesticide and soil matrix which inhibit a realistic analytical result.
- Analytical check on what? A dose verification? Soil extraction? Biodegradation in the soil? Please clarify first.
- It needs to be defined and explained how and why this should be necessary
- The appropriate guideline for understanding the fate and behaviour of the test substance in soil is OECD 307. Additionally, the use of artificial soil in the earthworm acute study would mean such information has little relevance to real world scenarios. However, verifying correct dosing by analysing the dosing solutions would add value to the study.
- In case studies are performed under GLP conditions and Certificates of Analysis are provided for each test substance, the amounts of substances or products applied do not need to be analytically verified. Moreover, analytical measurements in GLP soil laboratory tests are in many cases not meaningful due to multiple analytical difficulties when extracting from the soil matrix.
- Analytics will increase costs with limited benefit. There might be specific test thesis where analytic is helpful. Historically the old approach is sufficient and provides reliable information. Before an inclusion of analytics the use of the information should be discussed also in the context of a potential risk assessment. It should be kept in mind that the compulsory inclusion would lead to a huge repetition of studies unless there are considerations that old studies with no analytics are still valid

2. The amount of organic matter used should be decreased from 10 % to 5 %.

Rational: 10 % organic matter is not representative for many natural soils.

Furthermore, a high amount of organic matter likely reduces the bioavailability of the test chemical and therefore reflects a best case rather than a worst case scenario.

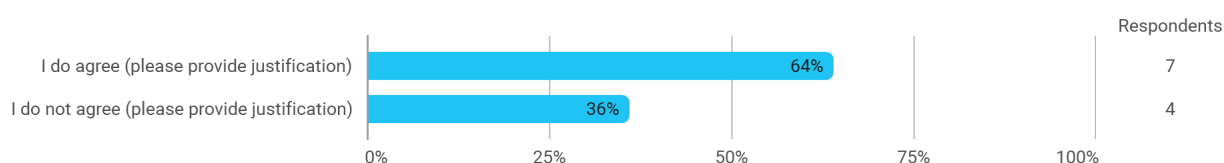


Disagreement justification:

- ▶ Eisenia would starve in case of lower organic matter content in soil and no additional food provision. In case of 5% OM food provision of manure is obligatory to avoid overarching starving effects.
- ▶ The species Eisenia does not live in agricultural soil with less organic content anyway. But the possibility to use 5% Peat should be added.
- ▶ Why should this be necessary? Why not leaving both options?
- ▶ Instead an option to use 5% as an alternative to 10% should be included, with guidance for when to use either option. Care should be taken that any changes to the test guideline do not invalidate previous studies, necessitating repeating significant numbers of animal studies.
- ▶ Eisenia would starve in case of lower organic matter content in soil and no additional food provision. In case of 5% OM, food provision of manure is obligatory to avoid overarching starving effects.
- ▶ The test is based on an artificial soil. The system was calibrated with 10% peat. In order to have the most flexible option it might be good to add a note (like in OECD 222) that a optional reduction of peat content may be done to address certain questions. the study. However there might be the option to add a note that studies with 5 % peat content are possible the
- ▶ The test is based on an artificial soil. The system was calibrated with 10% peat. In order to have the most flexible option, it might be good to add a note (like in OECD 222) that an optional reduction of peat content may be done to address certain questions.

3. The simple paper contact toxicity should be removed from the revised version of the TG.

Rational: This method is one of the oldest TG (1984). There is a need to update it in depth. The consistency with ISO 11268-1 should be improved.



Agreement justification:

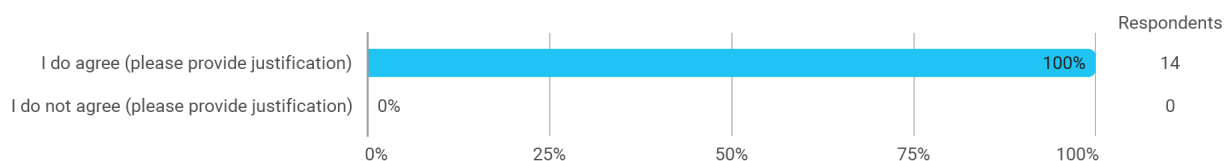
- ▶ Test not longer needed
- ▶ The paper filter test does not reflect the behaviour of the test item and test species in the soil
- ▶ Test not longer needed
- ▶ not used anymore in the past

Disagreement justification:

- ▶ In some cases, the paper test can be used as a pre-test option. Therefore, it should be kept.
- ▶ It is true that the paper test does not play a prominent role, however i do not see a need to remove it unless the reliability/reproducibility is questioned.

4. In addition, other reference chemical (e.g. boric acid) should be included.

Rational: This method is one of the oldest TG (1984). There is a need to update it in depth. The consistency with ISO 11268-1 should be improved

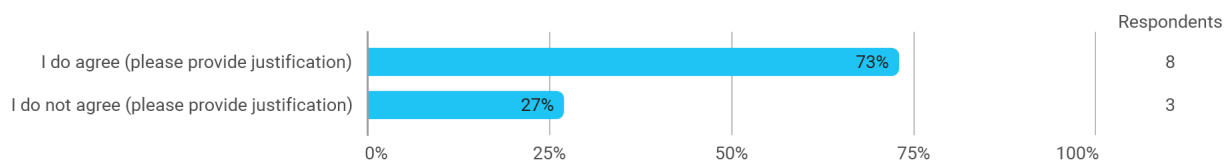


Agreement justification:

- ▶ Variety of reference item should be available
- ▶ Please delete boric acid due to worker safety reasons. The sentence should be kept. "In addition, other reference chemical should be included. "
- ▶ An isolated population lab can be more robust to one single reference item. Alternatives should be noted
- ▶ The reference item carbendazim is no longer available.
- ▶ However, dose of boric acid probably very high to investigate effects on mortality
- ▶ Please delete boric acid since it is not suitable due to worker safety reasons (CMR classification!). The sentence should be kept: "In addition, other reference chemicals should be included." A ring test should be performed with candidate substances which should preferably be organic substances exhibiting a medium LogP (e.g. 2).
- ▶ The reference item carbendazim is no longer available.
- ▶ another reference can be included. However the reference item should not be included in every test.
- ▶ Another reference can be included but the old reference should remain to not invalidate old studies. However the reference item should not be included in every test and should be tested occasionally as mentioned before in the guideline. It should be also kept in mind that boric acid has a very unfavourable hazard classification (repro Tox Cat 1B, may damage fertility/unborn child)

5. The possibility to use artificial soil prepared with 5% peat and natural soil should be included.

Rational: This method is one of the oldest TG (1984). There is a need to update it in depth. The consistency with ISO 11268-1 should be improved.



Agreement justification:

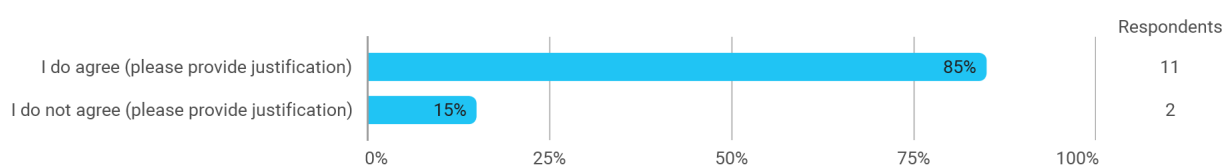
- ▶ The lower organic content would be more comparable to agricultural soils

- ▶ see point 2 above. It would be helpful to provide guidance on what natural soil properties would be considered acceptable and in what circumstances natural soils can be used.
- ▶ this should be included as option like it is done in OECD 222 as additional note
- ▶ This should not be included as mandatory but as optional as it is done in OECD 222 (as additional note)

Disagreement justification:

- ▶ Eisenia would starve in case of lower organic matter content in artificial or natural soil and no additional food provision. In case of 5% OM food provision of manure is obligatory to avoid overarching starving effects.
- ▶ natural soil make sense only in earthworm reproduction tests

6. Alternatives to sphagnum peat as a component of OECD artificial soil should be investigated.



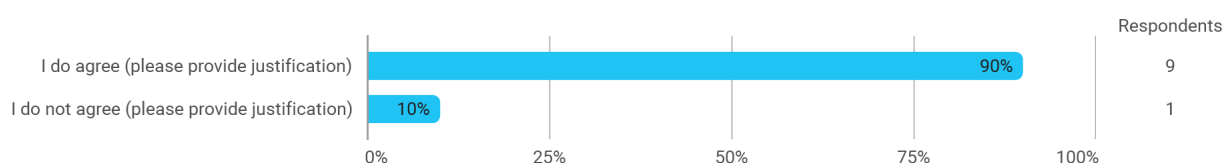
Agreement justification:

- ▶ If It seems to be a problem in the future to use phagnum peat, it would be good to look for alternatives
- ▶ Peat moss is not sustainable
- ▶ limited availability of peat
- ▶ sphagnum peat is very variable; to establish a more homogenous alternative would be helpful

Disagreement justification:

- ▶ Since the artificial OECD soil provided reliable and reproducible results, I do not see the big need for alternatives. Regardless a search for new alternative might be appropriate but would also need for additional feasibility/ring testing.

7. para B - Deionised water is added to give an overall moisture content of about 40 - 60 % of the WHCmax, and the medium is thoroughly mixed.



Agreement justification:

- ▶ As in the other soil organism TG, the moisture should be adjusted to approx. the half of the WHC

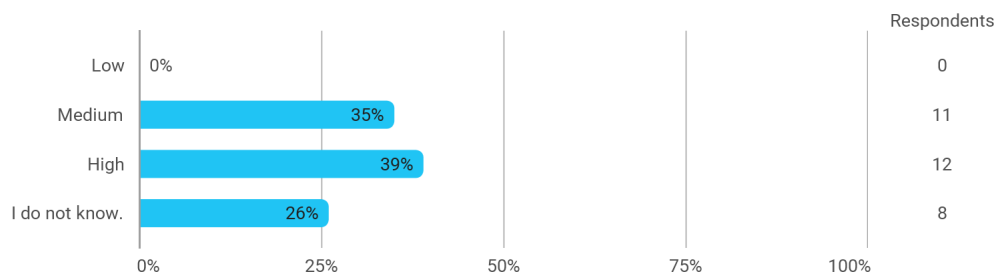
- Equal to other soil organism tests
- however wording should be aligned with ISO and other OECD guidelines

Disagreement justification:

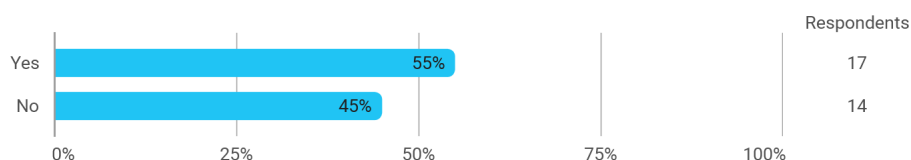
- Unclear what the intent of this comment is.

B.14 Test No. 216: Soil Microorganisms: Nitrogen Transformation Test

Relevance of revision

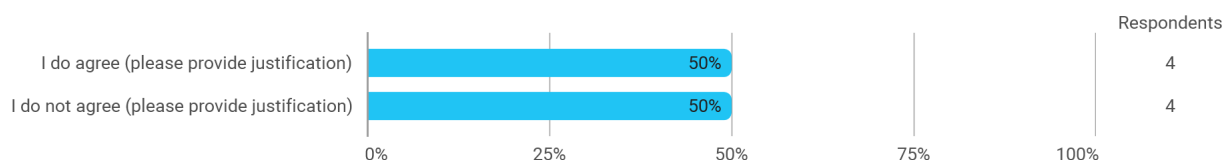


Interested to be part of an OECD Expert Group



The following suggestions have been collected:

1. **Verify the possibility to include the analytical verification (e.g., verification of the test item concentration in the treated soil) as for other tests. Include the exact formula for the calculation of the rate of nitrate because is not clear if the deviation should be calculated from 0 to 28 days or from 14 to 28 days (if no prolongation of the test is required). Is it required a toxic reference control for this test?**



Agreement justification:

- for verification of the test item concentration
- analytical verification should be considered as in other terrestrial tests

Disagreement justification:

- This suggestion is three separate points. Inclusion of analytical verification – disagree – it would only be appropriate to analytically confirm dosing e.g. by analysis of the stock solutions. This is in line with the requirements in Non-Target Terrestrial Plant studies (OECD 208 and OECD 227). This should not be used for understanding the behaviour of the test item as the appropriate guideline for understanding the fate of the test substance is

OECD 307. Formula for nitrate calculations - Agree – high importance. Clarification of the correct endpoint for use is required. The nitrate formation rate in the 0-28 day interval is the correct and valid interval for use, and this has been confirmed by the UBA via the OECD. This should be acknowledged in the guidance and the formula for calculating the nitrate formation rate should be provided. Requirement for a toxic reference –Agree - Medium importance. The inclusion of a toxic reference requirement in the guidance could add additional confidence in the sensitivity of the test system. Toxic reference controls are already conducted in many contract research organisations. It would be helpful if the guidance was updated to detail which toxic reference to use to enable standardisation. This would need to be backed by sufficient ring-testing and validation. Consideration of the frequency of testing would need to be considered and this could reflect the guidance given in OECD 222 (para 7)

- ▶ For agrochemicals: In case studies are conducted under GLP, there is no need to verify the test substance amount applied. The analytical concentration(s) in the test item is/are given in the Certificate of Analysis. The tested concentrations are given in the raw data, since the respective original weights are noted here. Because the samples are processed with a closed mixer after the corresponding solutions / mixtures have been added to natural soil, it is ensured that the test item quantity to be tested has reached the respective sample. Moreover, in general, a high logP of a test substance can lead to incorrect (i.e. too low) analytical values for the active ingredients, since they are difficult to be removed from soil matrix once they are in contact. The calculation of the transformation rate is a simple mathematical calculation (value divided by time difference). To determine an exact transformation rate from 0 to 28 days, an additional sample for the 0 to 28 days analysis would have to be included in the test system for all test concentrations. A calculation from 0 to 28 days makes no sense because of the storage of the samples. At day 0 the microbial population in the control can differ from the day 28 population. Reasons could be, among others, the storage of soil before test start (age of the used soil) or adaption of the microorganisms to the optimized humidity in the test system. Toxic reference tests make sense. They can be performed with a suitable reference substance (e.g. Dinoterb). A test conducted regularly (once per year) to determine the sensitivity of the test system is enough.
 - ▶ I partly agree. This is a multiple question. I disagree to analytical verification as it increase costs and provide very limited information for the overall assessment of the study. I agree that the calculation of the relevant endpoint (0-28d) should be described as relevant interval for risk assessment. A toxic reference control is not needed.
2. **para 6 - The suggested update is to clarify the evaluation of results for non-agrochemicals. Under paragraph 6, it stated that for non-agrochemicals, quantities of nitrate formed in treated and control samples are measured after 28 days. Paragraph 26 states that the quantities of nitrate in treated and control soil samples are determined and the ECx values are calculated. Paragraph 28 indicates that the data obtained on day 28 are used to determine the ECx, but that data from day 0 control samples can be used to report the initial nitrate level in soil. However, under paragraph 31, it is stated that for non-agrochemicals, the quantity of nitrate formed in each replicate is determined, and that the quantities of nitrate found in the treated samples after 28 days are compared to that found in the control. The text is not unequivocal; "formed" refers to a process and suggests that initial nitrate levels should be subtracted from the quantities found at day 28. This is however not mentioned. Therefore, it appears that for non-agrochemicals, the evaluation should be**

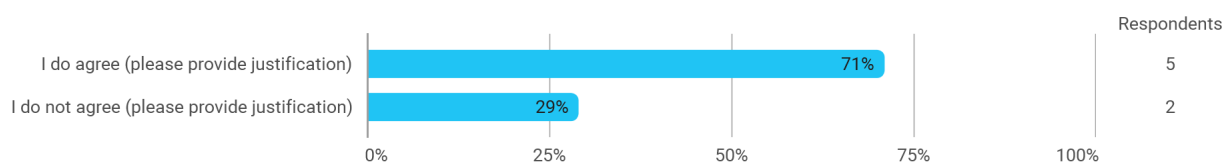
based on nitrate quantities measured at day 28 only, and initial nitrate quantities are not taken into account. Is this correct? If so, what is the reason for not considering initial nitrate levels? High nitrate levels at test initiation may reduce the effects seen during the test, especially if reduction of nitrification is not very large. For agrochemicals, the evaluation is based on formation rate (in consecutive intervals, even though not explicitly stated in the text) - and this includes initial nitrate levels by default. This approach seems better suited to detect small responses to chemical exposure. Please clarify what exactly should be done for non-agrochemicals and consider including initial nitrate levels in the assessment.

Rational: The text is not unequivocal and therefore subject to interpretation. Tests may be rejected if not conducted in accordance with interpretation of the evaluator, which can easily be prevented if the text is more clear.

para 6 - Add the following clarification to the 4th sentence (addition in bold Italic):

'The rate of nitrate formation in treated samples is compared with the rate in the controls, and the percent deviation of the treated from the control is calculated for the separate sample intervals (i.e., 0-7 d, 7-14 d, 14-28 d and further if applicable).'

Rational: Some applicants calculate the formation rate for the total study duration (0-28 d) or for intervals related to t= 0 (0-7 d, 0-14 d, 0-28 d). However, amongst risk assessor Member states the general agreement is that this is not what is intended with the test and that in fact only calculation per separate, subsequent time interval makes sense (if not, why would the samples at the intermediate time points be necessary?)



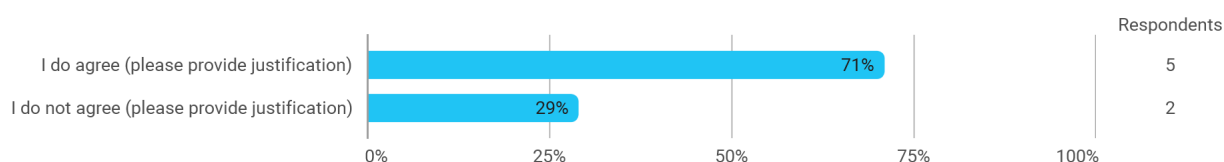
Agreement justification:

- ▶ to clarify the evaluation of the test results
- ▶ This refers to non-agrochemicals, but clarification of the correct endpoint for use is required, as discussed above in point 1
- ▶ to clarify the evaluation of the test results

Disagreement justification:

- ▶ There are several questions included in one point and several considerations that do not reflect the guideline. However it is true that a better wording would help. In general it is crucial to consider the historic considerations in the development of the guideline (e.g. Domsch (1983) and Gerber (1991)) where it became clear that the 0-28 day interval is the relevant endpoint for determining unacceptable effects. There is a general misunderstanding on the intermediate intervals. From a historic perspective of the guideline development (see Gerber 1991 and Domsch 1983) the goal was to differentiate reversible from irreversible or long term responses. As long term effects were considered as unacceptable the focus should be on the 0-28 day response as the relevant response for assessment purposes, while intermediate intervals are helpful to detect reversible responses which were not considered as unacceptable but were used to describe the course of the transient inhibition.

3. paras 6, 30, 31, 33 - In Paragraphs 6 (Principle of Test) and 30 (Data and Reporting; Data) it needs to be clearer which transformation rate should be reported as the endpoint. Parts of the Guideline allude that the rate over whole test period Day 0-28 transformation rate should be used but it does not categorically state this anywhere in the Guideline. There is the potential for mis-interpretation of the current guidance and transformation rates over 7/14 day intervals (i.e. Day 0 to Day 7; Day 7 to Day 14; Day 14 to Day 28), or transformation rates over incremental time periods of Day 0 to each sampling time point (i.e., Day 0 to Day 7, Day 0 to Day 14, and Day 0 to Day 28) could be used (There have previously been received comments suggesting the Day 0 to Day 7; Day 7 to Day 14; Day 14 to Day 28 rates should be used?). More clarity is given in Paragraphs 31 (Data and Reporting; Data) and 33 (Data and Reporting; Interpretation of results), where it appears the guideline is recommending that the Day 0-28 rate should be used/compared, but again it could be made much clearer. Rational: This revision is necessary as currently the CTGB and the UK CRD interpret the method for determining the nitrate formation rate differently to other regulatory authorities. This has led to the same study being considered to Pass the RA in one country and Fail the RA in another country. In our view the formation rate should be calculated (mg test item/kg soil d.w./day) for the whole test period (i.e. days 0-7, days 0-14, days 0-28). This would seem to best represent the data considering nitrogen source is added only at the study start and not at each sampling point.



Agreement justification:

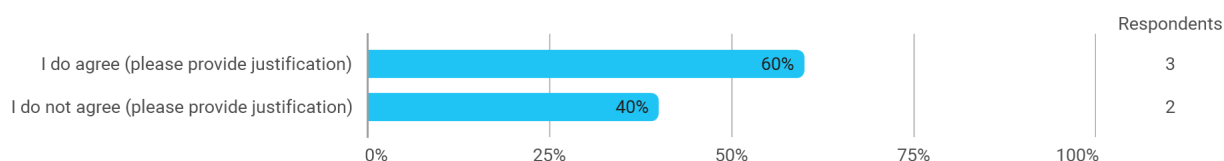
- It is currently unclear how the rates are to be calculated.
- It is important to better highlight the 0-28 d interval as the relevant result to determine unacceptable effects. Recap of the history of the guideline might be helpful to confirm the 0-28 d interval (and not the intermediate intervals) as relevant

Disagreement justification:

- Point 2 above is given twice. This response is in relation to the second point two in the original list: The 0-28 day period is the correct and valid interval for use, and this has been confirmed by the UBA via the OECD. This should be acknowledged in the guidance.

4. paras 6, 27, 26 - Clarification on the study duration when the test is extended beyond 28 days.

Rational: In Paragraph 6 (Principle of Test), Paragraph 26 (Performance of Test; Test conditions and duration) and Paragraph 27 (Performance of Test; Soil sampling schedule) it states the maximum study duration is 100 days, but this does not tally with Paragraph 27 it states that sampling timepoints beyond Day 28 should be carried out at 14 day intervals (which equates to Day 42, 56, 70, 84 and 98). If effects are seen at Day 98 should additional sampling be undertaken on Day 100?

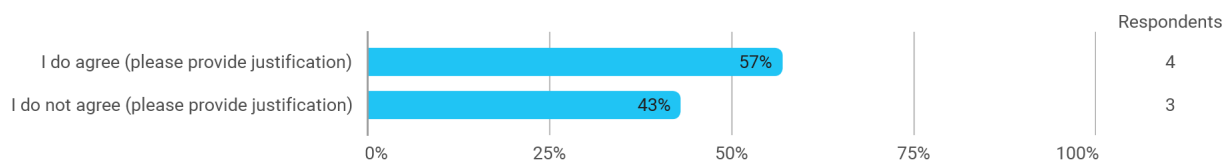


Agreement justification:

- ▶ Given the numbering error, this comment is in response to point 3 above: Clarification of the correct endpoint for use is required. The nitrate formation rate in the 0-28 day interval is the correct and valid interval for use, and this has been confirmed by the UBA via the OECD. This should be acknowledged in the guidance and the formula for calculating the nitrate formation rate should be provided.
- ▶ see comments before

Disagreement justification:

- ▶ The guideline clearly states in para 26: The minimum duration of tests is 28 days. If agrochemicals are tested, the rates of nitrate formation in treated and control samples are compared. If these differ by more than 25 % on day 28, the test is continued until a difference equal to or less than 25 % is obtained, or for a maximum of 100 days, whichever is shorter.
5. **para 7 - In the "validity of the test" section (7. on page 2) of the current test guideline, it should be clarified in the last sentence, that the coefficient of variation should be used. Current text: "Therefore, the variation between replicate control samples should be less than +/- 15%." Proposed new text (change in bold): "Therefore, for nitrate concentration the coefficient of variation between replicate control samples should be less than +/- 15%." Additional question: why is here used the nitrate concentration and not the nitrogen transformation rate?**
- Rational: Clarification is missing, which leads to variation in the analysis of the results.**

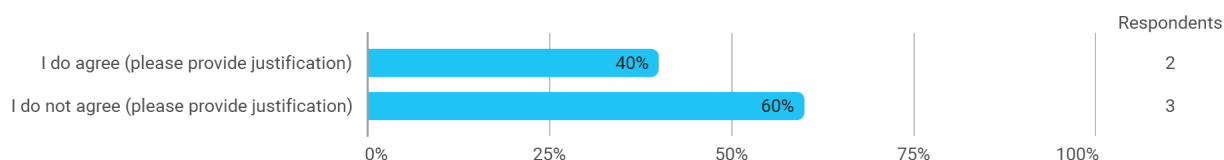


Agreement justification:

- ▶ Given the numbering error this is in relation to point 4 above: the guidance is unclear. Paragraph 26 "...If these differ by more than 25% on day 28, the test is continued until a difference equal to or less than 25% is obtained, or for a maximum of 100 days." Paragraph 27 says the test is extended in 14-day intervals but extending in 14-day increments results in a 98-day test. This wording could be clarified to make the testing timepoints required clearer.
- ▶ I do agree. About the second question: Nitrate is the measured compound, and therefore it is correct to mention the nitrate concentration. The nitrogen transformation rate is a calculated value.
- ▶ no further comment needed

Disagreement justification:

- ▶ the experience shows that it is very difficult to fulfil the validity criterion of CV 15 % based on the nitrogen transformation rate
6. **para 20 - In an additional paragraph, analytical measurements should be described. Analytical measurements should be performed for all substances, covering the concentration range. Here, either the concentration in total soil or the concentration in pore water should be determined. Following the EFSA Scientific Opinion on in-soil organisms (2017), “it is recommended that the concentrations of active substances in soils is measured more than twice (see also the EFSA DegT50 EFSA Guidance), depending on the degradation of the active substances and the length of the test.”**
Rational: According to EFSA Scientific Opinion on in-soil organisms (2017, chapter 7.11.3), analytical measurements for tests on in-soil organisms are requested. In this Scientific Opinion, it is stated that “During laboratory handling procedures of the spiked soils, however, possible losses of the pesticides due to volatilization, degradation, and sorption to, e.g. the glass matrix of vessels used, may occur.”
Therefore, the actual concentration of the tested substance needs to be determined in order to correctly link exposure and effects. Moreover, analytical measurements in test systems are included in all aquatic studies. Terrestrial studies should be analyzed in the same way, otherwise ‘real’ exposure cannot be described and an extrapolation from lab to field is questionable.



Agreement justification:

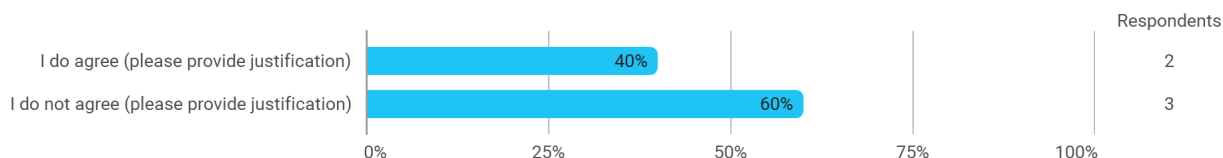
- ▶ Given the numbering error this is in relation to point 5 above: The coefficient of variation is already commonly reported, but clarification of the requirement would be helpful. Likewise, it should be clarified it this requirement relates to the nitrate concentration or nitrate formation rate.

Disagreement justification:

- ▶ In case studies are performed under GLP conditions and Certificates of Analysis are provided for each test substance, the amounts of substances or products applied do not need to be analytically verified. Moreover, analytical measurements in GLP soil laboratory tests are in many cases not meaningful due to multiple analytical difficulties when extracting from the soil matrix.
 - ▶ An analytical verification would significantly increase costs and with very limited additional information. The study is conducted under GLP considerations which provide a sufficient certainty on the correct exposure regime in the soil. If analytics is included, would it lead to an invalidation of all previous studies?
7. **para 26 - Add the following clarification to the 2nd sentence (addition in bold Italic):**
‘If agrochemicals are tested, the rates of nitrate formation in treated and control

samples are compared for the separate sample intervals (i.e., 0-7 d, 7-14 d, 14-28 d and further if applicable).'

Rational: Some applicants calculate the formation rate for the total study duration (0-28 d) or for intervals related to $t=0$ (0-7 d, 0-14 d, 0-28 d). However, amongst risk assessor Member states the general agreement is that this is not what is intended with the test and that in fact only calculation per separate, subsequent time interval makes sense (if not, why would the samples at the intermediate time points be necessary?)



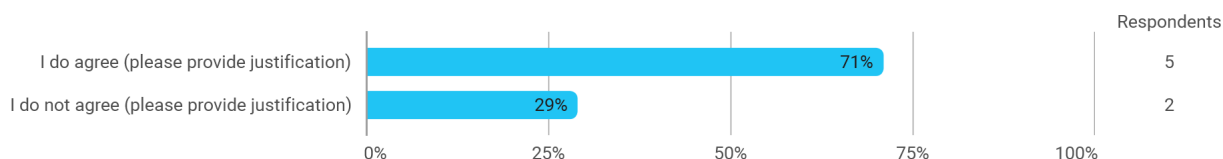
Disagreement justification:

- ▶ Given the numbering error this is in relation to point 6 above: The appropriate guideline for understanding the fate of the test substance in soil is OECD 307. A more appropriate approach is to analytically confirm dosing e.g. by analysis of the stock solutions. This is in line with the requirements in Non-Target Terrestrial Plant studies (OECD 208 and OECD 227)
- ▶ This is not the correct interpretation of the goal of the guideline. The goal of the guideline is to detect irreversible long term effects. Historically an interval of 0-28 days was considered as appropriate. Please refer to the workshops and Symposia where the guideline was developed. (e.g. Domsch (1983) and Gerber (1991))

- 8. para 30 - The quantities of nitrate formed are expressed in mg nitrate/kg dry weight soil/day. The duration applied for the calculation of the respective nitrate formation rate is always from test start (day 0) until the actual date. OR The duration applied for the calculation of the respective nitrate formation rate is always between two successive dates. The nitrate formation rate in each treatment is compared with that in the control, and the percent deviation from the control is calculated.**

Rational: The guideline does not explicit say, which measurement results (from which sampling dates) should be applied for the nitrate transformation rate calculation in tests with agrochemicals. The rates should be compared for all sampling dates.

However, should the rates from e.g. day 0 to day 7, from day 7 to day 14 and from day 14 to day 28 be compared, or from day 0 to day 7, from day 0 to day 14 and from day 0 to day 28? Both interpretations fit with the guideline wording. But there are huge effects on the rates them selves and the consequences. Effects in longer lasting periods (prolonged periods) remain longer than in successive (periods).



Agreement justification:

- ▶ clarification is necessary

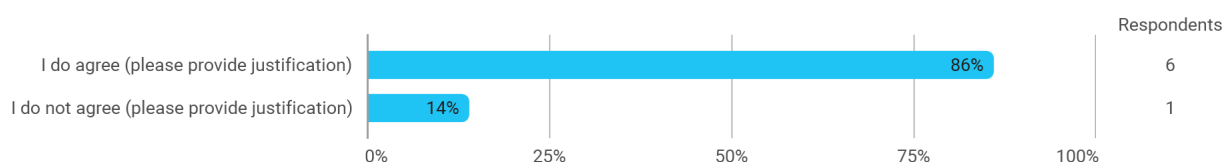
- ▶ I do agree with the first part. The duration applied for the calculation of the respective nitrate formation rate is always between two successive dates. The nitrate formation rate in each treatment is compared with that in the control, and the percent deviation from the control is calculated.
- ▶ The goal of the guideline is to detect irreversible long term effects. Historically an interval of 0-28 days was considered as appropriate. Hence intermediate intervals between successive dates is not correct. Please refer to the workshops and Symposia where the guideline was developed. (e.g. Domsch (1983) and Gerber (1991))

Disagreement justification:

- ▶ Given the numbering error this is in relation to point 7 above: The nitrate formation rate in the 0-28 day interval is the correct and valid interval for use, and this has been confirmed by the UBA via the OECD. This should be acknowledged in the guidance and the formula for calculating the nitrate formation rate should be provided.

9. **para 30 - In the test guideline there should be clearly stated/defined, whether the time interval that is used for the effect calculation (i.e. nitrogen transformation rate) shall be section-by-section (e.g. 0-7, 7-14, 14-28 days etc.) or whole period (e.g. 0-28 days). Currently this is not defined, which leads to an unclear situation for the result interpretation. (Is only one of the two options applicable or both? Which to use if both can be used?).**

Rational: The unclear situation prevents a harmonised presentation and analyses of the data derived from this guideline. Furthermore, the different methods of data analysis described above lead to different results. Hence, the guideline would benefit from the proposed revisions.



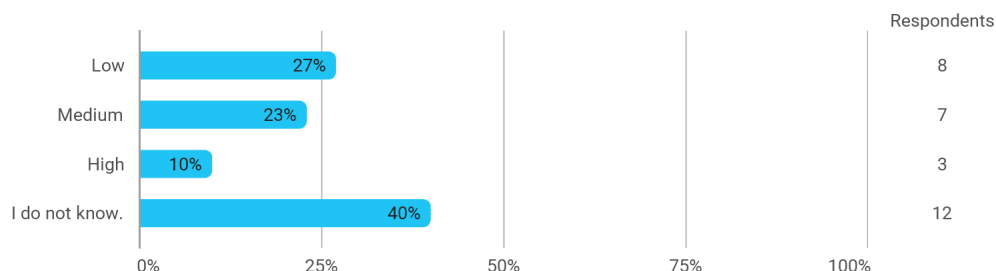
Agreement justification:

- ▶ clarification is necessary
- ▶ Given the numbering error this is in relation to points 8 and 9 above: Clarification of the correct endpoint for use is required. The nitrate formation rate in the 0-28 day interval is the correct and valid interval for use, and this has been confirmed by the UBA via the OECD. This should be acknowledged in the guidance and the formula for calculating the nitrate formation rate should be provided
- ▶ I do agree with the first part. The nitrate formation rate in each treatment is compared with that in the Control (section-by-section (e.g. 0-7, 7-14, 14-28 days etc.).
- ▶ clarification is necessary
- ▶ I agree that it should be better highlighted that the 0-28 day interval is the relevant interval to detect unacceptable effects, while the intermediate intervals are determined to give a better understanding on the short term formation but that it is not considered as relevant

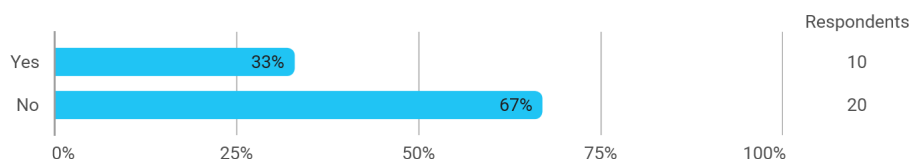
endpoint. A recap of the history of the guideline is helpful to differentiate the different considerations.

B.15 Test No. 220: Enchytraeid Reproduction Test

Relevance of revision



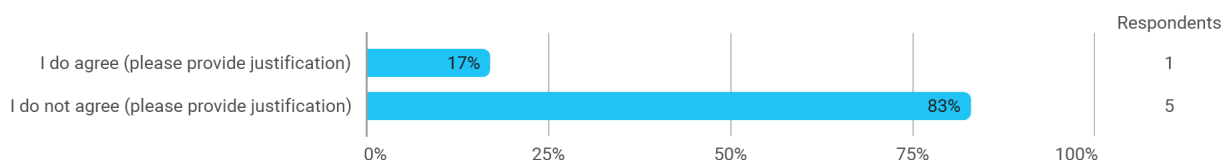
Interested to be part of an OECD Expert Group



The following suggestions have been collected:

1. It might be considered to make the requirement for analytical check for the behaviour of the chemical compulsory. The comment might be relevant for all tests using artificial soil.

Rational: It is mandatory for aquatic organisms (even for acute tests), including sediment dwellers. Indeed, the dissipation of the chemical largely influences the effect, therefore the endpoint get from the test (which is the main aim). The dissipation rate can considerable different in the test media than in the field. Therefore, the use of the endpoints to compare the toxicity of the different chemicals might be a questionable practice. What the endpoint actually express is a combination of the toxicity and the dissipation rate of the chemical in the test media.



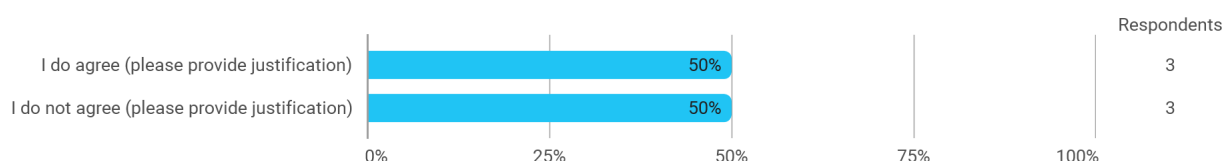
Disagreement justification:

- Analytical check on what? Dose verification, soil extraction, biodegradation? Please clarify first
- The appropriate guideline for understanding the fate and behaviour of the test substance in soil is OECD 307. Additionally, the use of artificial soil in the Enchytraeid reproduction study would mean such information has little relevance to real world scenarios. However, verifying correct dosing by analysing the dosing solutions would add value to the study.

- In case studies are performed under GLP conditions and Certificates of Analysis are provided for each test substance, the amounts of substances or products applied do not need to be analytically verified. Moreover, analytical measurements in GLP soil laboratory tests are in many cases not meaningful due to multiple analytical difficulties when extracting from the soil matrix.
- Analytics will increase costs with limited benefit. There might be specific test thesis where analytic is helpful. Historically the old approach is sufficient and provides reliable information. Before an inclusion of analytics the use of the information should be discussed also in the context of a potential risk assessment. It should be kept in mind that the compulsory inclusion would lead to a huge repetition of studies unless there are considerations that old studies with no analytics are still valid

2. para 11 - An alternative to sphagnum peat as source of organic matter should be provided.

Rational: The OECD artificial soil requires 10 % organic matter. It is recommended to use sphagnum peat. However, sphagnum peat is not available worldwide and its use is also questionable for ecological reasons.



Agreement justification:

- Peat moss is not sustainable

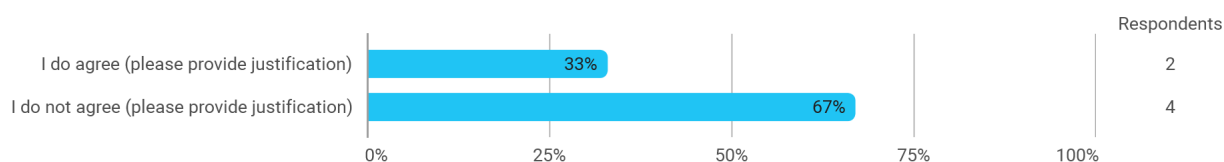
Disagreement justification:

- An alternative to sphagnum peat as source of organic matter can only be provided after a feasible validation process.
- since the artificial OECD soil provided reliable and reproducible results, I do not see the need for alternatives. It would also need for additional ring testing.

3. para 11 - The amount of organic matter used should be decreased from 10 % to 5 %.

Rational: 10 % organic matter is not representative for many natural soils.

Furthermore, a high amount of organic matter likely reduces the bioavailability of the test chemical and therefore reflects a best case rather than a worst case scenario.



Agreement justification:

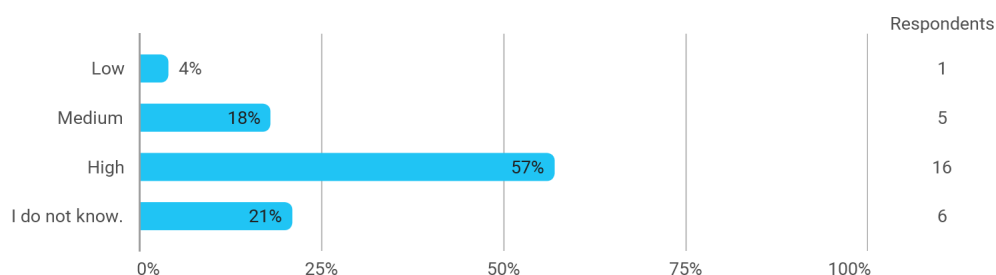
- Enchytraeids are abundant in agricultural soils with less organic carbon content

Disagreement justification:

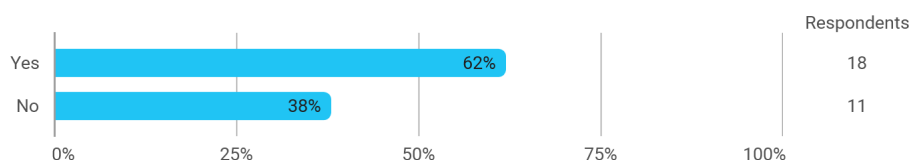
- ▶ Instead, options for 10% and 5% should be included, with guidance on when each is acceptable. Care should be taken that any changes to the test guideline do not invalidate previous studies, necessitating repeating significant numbers of animal studies.
- ▶ Enchytraeids would starve in case of lower organic matter content in soil and no additional food provision. In case of 5% OM, food provision of manure is obligatory to avoid overarching starving effects.
- ▶ However, a reduction can be mentioned as optional as it is done in OECD 222

B.16 Test No. 222: Earthworm Reproduction Test (*Eisenia fetida*/*Eisenia andrei*)

Relevance of revision

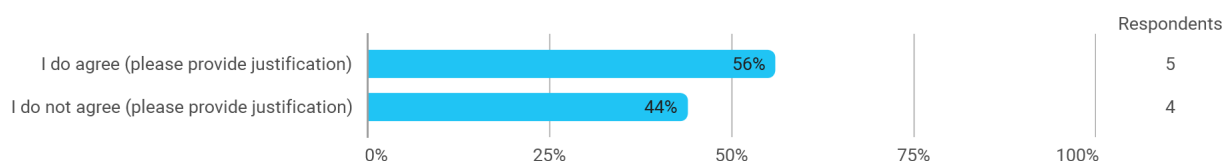


Interested to be part of an OECD Expert Group



The following suggestions have been collected:

1. **An alternative, optional test design should be included to allow data generation for parameterisation of toxicokinetic-toxicodynamic models (e.g., GUTS and DEB-TKTD). This alternative design should include measurements of organism size, mortality, and reproductive output at intermediate time points (at least 4). This can be achieved by destructive sampling of a test vessels. Replication at each sampling time is not necessary for parameterisation of toxicokinetic-toxicodynamic. Instead, replicate test vessels can be re-distributed for sampling at intermediate time points.**
Rational: Toxicokinetic-toxicodynamic modelling, especially when combined with population modelling, can help to quantify and reduce uncertainties in environmental risk assessment by providing improved integration of exposure and effects and by incorporating different pesticide application scenarios and greater ecological realism.
References: Forbes VE, Agatz A, Ashauer R, Butt KR, Capowicz Y, Duquesne S, Ernst G, Focks A, Gergs A, Hodson ME et al. 2021. Mechanistic effect modeling of earthworms in the context of pesticide risk assessment: Synthesis of the foresee workshop. Integrated Environmental Assessment and Management. 17(2):352-363.; SCENIHR (Scientific Committee on Emerging and Newly Identified Health Risks), SCHER (Scientific Committee on Health and Environmental Risks), SCCS (Scientific Committee on Consumer Safety). 2012. Addressing the new challenges for risk assessment. Brussels: European Commission.



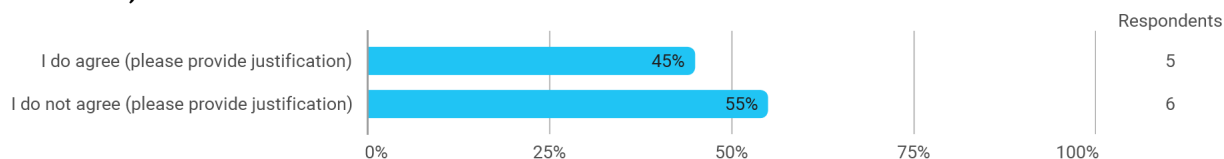
Agreement justification:

- Modelling can help generating data as experimental data sometimes cannot answer all questions in risk assessment
- If needed for toxicokinetic-toxicodynamic models
- The suggested updates to these guidelines will enable more realistic and ecologically relevant environmental risk assessments
- However further detailed considerations need to be done to define the needed parameters. In order to not overcomplicate the guideline it might be an option to include the alternative design in an appendix.

Disagreement justification:

- see bioaccumulation guideline
 - I do not agree to include this as a mandatory test design in the main chapters. Modelling is not an useful tool for all substances due to low toxicity of many substances. It must be included as Appendix (optional test design) with the following suggestions: Ideally, an optional test design would be included for targeted TKTD experiments. For earthworms, this should rather include biomass instead of size. The optional test design should include an earthworm cocoon test, as described by Rakel et al (in submission) “Earthworm cocoon test for TKTD Modelling in soil risk assessment of chemicals” and as presented at SETAC Europe 2021 by Rakel et al “Alive and kicking? – Earthworm cocoon test for TKTD Modelling in soil risk assessment of chemicals”. Details can be provided upon request.
 - Due to high variability and uncertainty associated with the 3-dimensional patterns of toxicant degradation and organism movement in soil (“patchiness”), exposure profiles for soil organisms can be predicted only with high uncertainty. This is already a key issue for Tier 1 experimental testing. Therefore, the use of TKTD models to predict effects on soil organisms in time and space as in the aquatic Tier 2C - where organism exposure in the water column or at the sediment surface is potentially more homogeneous in space- is highly questionable. Developing reliable TKTD models for predicting effects on soil organisms and setting up meaningful scenarios for TKTD modelling in order to identify a realistic worst-case would require huge (even unrealistic) efforts in characterizing the toxicant degradation and the test organism’s behaviour; the species tested should also be suitable to represent of potential toxicologically sensitive and ecologically vulnerable species under realistic field conditions. However, TKTD models could be more successfully applied to retrospectively analyse and understand differences in effects on soil organisms observed across different studies, e. g. across soils having different characteristics. In general, we agree that analytical measurements should be performed in laboratory studies in order to better link exposure to effects. In this regard, residues should be analyzed at different exposure levels in the soil, soil pore water and organisms at the beginning, during and at the end of the study.
2. **para 3 - Suggestion to bring in line with OECD 232, where it is more explicit in this section that mortality is also a relevant endpoint of the study.**

Rational: For OECD 222 and 226 it sometimes is a point of discussion whether mortality should be determined as endpoint for the study or not. In addition, there is a discrepancy between OECD 222 + 226 on the one hand, and OECD 232 on the other hand, as in OECD 232 it seems clear that mortality should be used as well for endpoint derivation, whilst in OECD 222 and 226 this is not clear.



Agreement justification:

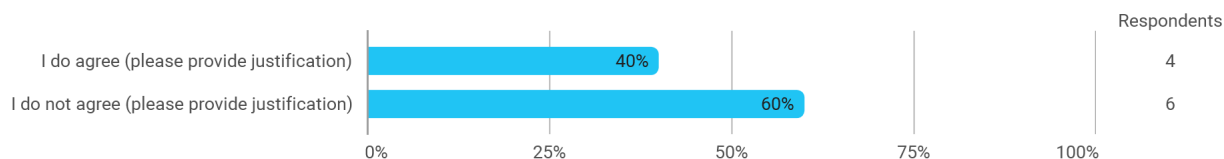
- ▶ Mortality is one relevant parameter and should be highlighted as well as the reproduction
- ▶ Mortality is also a relevant endpoint
- ▶ Mortality should continue to be assessed and reported, however changes to paragraph 3 to make mortality a required endpoint would necessitate changes throughout the guideline, which should be undertaken with caution. The focus of this study should remain on sub-chronic effects, particularly reproduction. If the test guideline is amended to make mortality a required endpoint, there should also be an acknowledgement that the maximum application rate tested should be focused around the expected sub-chronic toxicity range and not the mortality range I.e. mortality should be a nice to have rather than a need to have. The primary emphasis of this study design should continue to be reproduction.
- ▶ Agreement since mortality is definitely a relevant endpoint according to paragraph 42.

Disagreement justification:

- ▶ The reduction of reproduction should not be based on mortality. if possible, no mortality should occur up to the highest test concentration.
- ▶ this is a reproduction test; ideally no mortality should be observed in all concentrations tested
- ▶ Reproduction is the distinctly more sensitive endpoint. To assess the mortality, the OECD 207 is the appropriate test design to assess mortality and derive a meaningful LC value. NOEC based on mortality is assessed in the OECD 222 as well.
- ▶ An aligned terminology of 232 and 222 and 226 regarding mortality and reproduction data is appropriate. However, as reproductive data (number of juvenile) will include lethal mortality aspects (only survived animals will reproduce) the clear focus should be set to the reproduction data.

3. **para 5 - This is a more general question for soil ecotoxicology studies: the data requirements for active substances section 4 (analytical methods) states that: "Methods shall be submitted, with a full description, for the determination of non-isotope-labelled residues in all areas of the dossier, as set out in detail in the following points: in soil, water, sediment, feed and any additional matrices used in support of ecotoxicology studies;" However in practice, analytical methods for soil ecotoxicology studies are almost never provided, since it is not a direct requirement in the guideline. It is still defined as "optional": For volatile, unstable or readily degrading**

substances (e.g., data generated from a TG 307 study may be considered), or where there is otherwise uncertainty in maintaining the nominal soil concentration, analytical measurements of the exposure concentrations at the beginning, during and at the end of the test should be considered. Would it be possible to clarify that analytical verification of test concentrations in soil is required in the guideline?



Agreement justification:

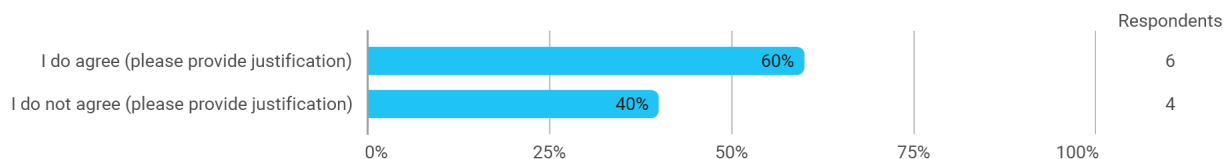
- ▶ As option to confirm exposure over time.
- ▶ Agreement. Instead, analytical measurements should be performed for all substances, covering the concentration range. Here, the concentration in total soil as well as the concentration in pore water should be determined. Following the EFSA Scientific Opinion on in-soil organisms (2017), "it is recommended that the concentrations of active substances in soils is measured more than twice (see also the EFSA DegT50 Guidance), depending on the degradation of the active substances and the length of the test."

Disagreement justification:

- ▶ If the test item is readily biodegradable etc. what is the consequence for test system? Applying the test item at the middle of the test again? Clarification of the analytical methods (dose verification, soil extraction...) should be made prior to renewal of the guideline
 - ▶ It needs to be defined and explained how and why this should be necessary
 - ▶ The appropriate guideline for understanding the fate and behaviour of the test substance in soil is OECD 307. Additionally, the use of artificial soil in the earthworm reproduction study would mean such information has little relevance to real world scenarios. A more appropriate approach is to analytically confirm dosing e.g. by analysis of the test solutions.
 - ▶ In case studies are performed under GLP conditions and Certificates of Analysis are provided for each test substance, the amounts of substances or products applied do not need to be analytically verified. Moreover, analytical measurements in GLP soil laboratory tests are in many cases not meaningful due to multiple analytical difficulties when extracting from the soil matrix.
 - ▶ An inclusion of analytics should be justified by reasonable means. Studies where performed without analytics without any significant issues. Analytics would again increase the costs and complexity. There might be specific questions where analytics is justified, however it should remain as an optional point. Before an inclusion of analytics, the use of the information should be discussed also in the context of a potential risk assessment. It should be kept in mind that the compulsory inclusion would lead to a huge repetition of studies unless there are considerations that old studies with no analytics are still valid
- 4. para 5 - Further guidance on follow up of the last sentence of this paragraph ('For volatile (...) considered.') is considered necessary. I.e.: criteria to be used for deciding when a substance is considered 'unstable' (e.g. a DT50-value) when measurements are performed, criteria for when to use nominal or measured concentrations**

(probably the 80% criterium) how to express measured concentrations (i.e. geomean, twa) when no measurements are performed while considered necessary based on the above, what are the consequences for the validity of the study.

Rational: The 'strictness' of the last sentence of this paragraph ('For volatile (....) considered.'), is not very clear and it leaves open several questions.



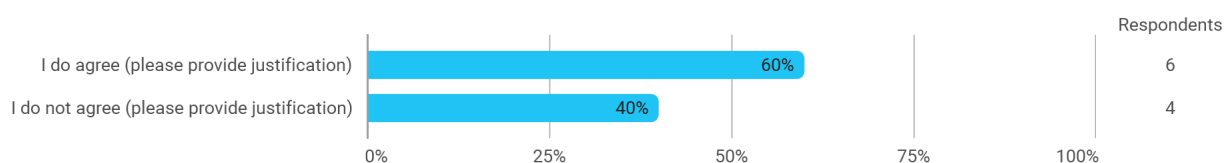
Agreement justification:

- ▶ Never been confronted with this problem before, but it should be clarified
- ▶ Further guidance in helpful
- ▶ specification is needed
- ▶ Agreement. This should be clarified because currently this paragraph leads to the fact that analytical measurements are never performed.

Disagreement justification:

- ▶ The determination of a mean measured soil concentration or a twa is not appropriate for this study type. Not least because it is not possible to design a study type that allows concentrations to be maintained e.g. a flow through system. The risk assessment is conducted by comparing the peak initial concentration from the exposure models with the effects that are seen at that peak concentration. The current test design reflects the risk assessment requirements
- ▶ In case studies are performed under GLP conditions and Certificates of Analysis are provided for each test substance, the amounts of substances or products applied do not need to be analytically verified. Moreover, analytical measurements in GLP soil laboratory tests are in many cases not meaningful due to multiple analytical difficulties when extracting from the soil matrix.
- ▶ It should be kept in mind that detection of residues in soil has different complexity than for example in the aquatics where the 80% criteria is used. In general analytics in soil lab studies will lead to a lot of challenges regarding the interpretation.

5. para 7 - Delete the last sentence



Agreement justification:

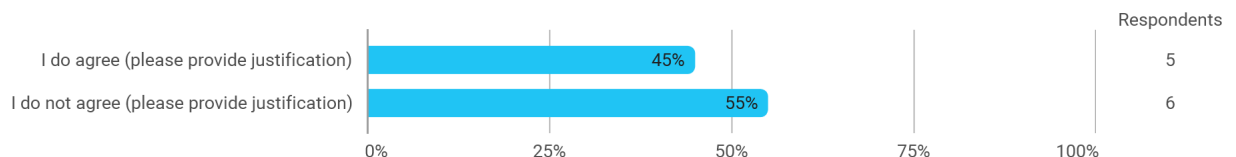
- ▶ In practical, a toxic standard is not included in the test series. Toxic reference test are conducted once or twice a year to show adequate sensitivity of the population used
- ▶ If not needed anymore

- ▶ unusual procedure
- ▶ unusual procedure
- ▶ deletion is reasonable

Disagreement justification:

- ▶ The last sentence seems a reasonable approach if including a positive control in the test series alongside having regular full series tests.
- ▶ The sentence should not be removed but modified. A toxic reference test performed once a year makes sense. It is however NOT necessary to include a positive toxic reference substance in each test. There should be free choice for the test design (i.e. concentrations, replicates, provided the results are meaningful) for the toxic reference test. An acceptable range for the sensitivity should be given. In case a new toxic reference is suggested it should be ring-tested.
- ▶ Disagreement since I do not understand the reasoning for deleting the last sentence “If a positive toxic standard is included in the test series, one concentration is used and the number of replicates should be the same as that in the controls.”

6. para 7 - Delete Carbendazim & Benomyl and use instead boric acid



Agreement justification:

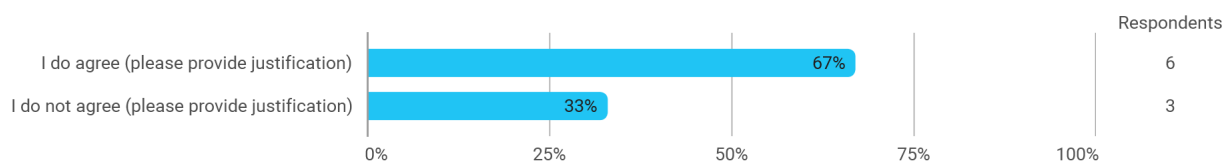
- ▶ limited availability of carbendazim
- ▶ boric acid is proven to be a substitute for carbendazim and benomyl
- ▶ limited availability of carbendazim
- ▶ Agreement that Carbendazim & Benomyl are not to be used as reference substances anymore since Carbendazim is not approved in the EU anymore.

Disagreement justification:

- ▶ Carbendazim is a well known toxic reference. Boric acid should be added.
- ▶ Add boric acid
- ▶ Boric acid should be included as an alternative reference item option. Care should be taken that any changes to the test guideline do not invalidate previous studies, necessitating repeating significant numbers of animal studies
- ▶ Boric acid is not a suitable substance due to its high toxicity (worker safety). It is classified as CMR.
- ▶ The old toxic reference should still be mentioned as they provide reliable results. However new references can be mentioned as preferred. It should be also kept in mind that boric acid

has a very unfavourable hazard classification (repro Tox Cat 1B, may damage fertility/unborn child). So the advantages of boric acid should be discussed

7. para 7 - According to ISO effect for boric acid should be 400 - 600 mg boric acid/kg soil. A proposal for test concentration for boric acid, based on our and other lab experiences, could be 200 - 400 mg boric acid / kg soil



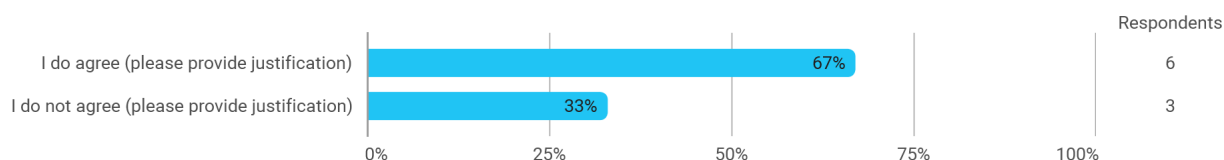
Agreement justification:

- ▶ The effect range should be adjusted to practical data and experience
- ▶ Add information for boric acid in addition to there reference items
- ▶ subject to appropriate ring testing/data gathering to confirm toxicity range.
- ▶ Maybe, not a range of concentrations but a range of EC50 values should be given
- ▶ Thank you for your valuable input and trials, that is very helpful and results could potentially be used for the guideline

Disagreement justification:

- ▶ I do not agree
- ▶ Boric acid is not a suitable substance due to its high toxicity (worker safety). It is classified as CMR.
- ▶ There should be a reliable data set provided to confirm this before it is included in the guideline . Until then it would be more appropriate to align with ISO

8. para 8 - The mean number of juveniles per replicate should be at least 30 at the end of the test (see Folsomia guideline OECD 232)



Agreement justification:

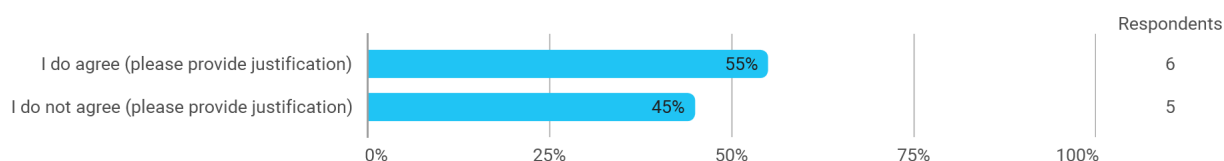
- ▶ TG for the same ecological containment should be harmonised
- ▶ should be accessible
- ▶ However it is not clear why this is mentioned here as it is already implemented in OECD 222 (see Par 8 validity criteria)

Disagreement justification:

- ▶ This change is unnecessary. Current wording is appropriate

- The mean number of juveniles per vessel should preferably be 60 at the end of the test for *Eisenia* species in artificial soil with sufficient food. In case other more slowly and less reproductive earthworm species are tested, lower numbers are acceptable.

9. para 9 - The containers should have a cross-sectional area of approximately 200 cm² so that a moist substrate depth of about 5-6 cm is achieved when 500 to 600 g dry mass of substrate is added.



Agreement justification:

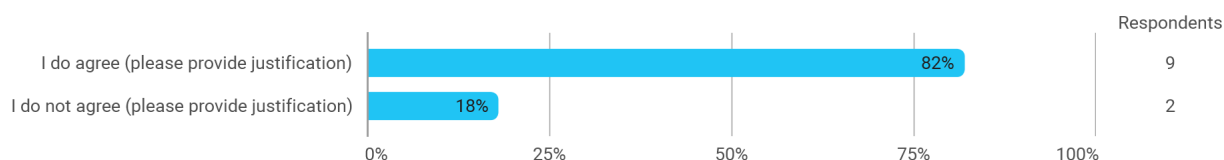
- What is the question here? The description is good for common plastic vessel in 18x14x6 cm
- already standard
- this is the text that is already in the guideline
- We appreciate this suggestion and support the harmonization of test containers.

Disagreement justification:

- I cannot see a difference
- Change not required, this is the current wording.
- This is the original text of the guideline. To fulfil these requirements is not possible. The proportions within this text must be adapted to each other.
- For *Eisenia*, the depth of soil is not relevant in this test. Therefore, the size of the containers should be kept flexible. Generally, a large surface allows for a precise estimation of food consumption. Especially for other earthworm species, different dimensions and amounts of soil should be possible. Justification should then be provided.

10. para 11 - Alternatives to sphagnum peat as a component of OECD artificial soil should be investigated

Rational: The OECD artificial soil requires 10 % organic matter. It is recommended to use sphagnum peat. However, sphagnum peat is not available worldwide and its use is also questionable for ecological reasons.



Agreement justification:

- Peat moss is not sustainable
- according to ISO coco peat
- In the interests of sustainability alternatives to peat should be actively investigated.

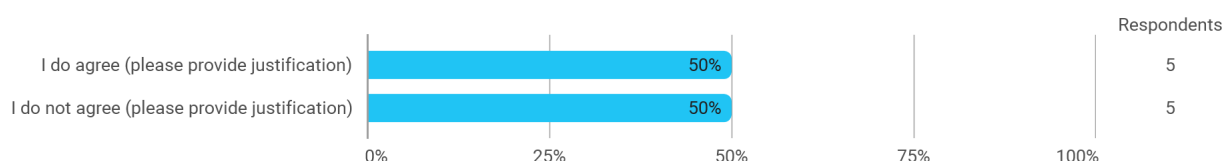
- ▶ sphagnum peat is very variable
- ▶ according to ISO coco peat
- ▶ The extraction of peat contributes directly to climate change. In contrast, using alternatives would contribute to species conservation and climate protection. In the future, the extraction of peat could also be prohibited by law for climate protection reasons. Therefore, it is to be supported to look for alternatives soon.

Disagreement justification:

- ▶ since the artificial OECD soil provided reliable and reproducible results, I do not see the need for alternatives. It would also need for additional feasibility/ring testing.

11. para 11 - The following sentence should be changed with respect to the percent sphagnum peat to the following: “5 percent sphagnum peat (as close to pH 5.5 to 6.0 as possible, no visible plant remains, finely ground, dried to measured moisture content)”.

Rational: Following the “Scientific Opinion addressing the state of the science on risk assessment of plant protection products for in-soil organisms”, EFSA (2017), there is a need for a more standardized test design, e.g. with respect to the peat concentration. In the OECD guideline, 10 % sphagnum peat is still the standard concentration for the description of the peat content whereby in most studies, only 5 % sphagnum peat content in test soils is used. In addition to having a more standardized test design for in-soil organisms, a peat content of 5 % represents natural agricultural soils showing a carbon content of 1-5 % much more realistic. Therefore, questions of sorption as well as bioavailability of test chemicals are reflected much better in test systems which reduced organic matter content.



Agreement justification:

- ▶ For the reproduction test the soil similar to folsomia and hypoaspis should be used
- ▶ already standard
- ▶ We support this suggestion as 5 % organic carbon content is closer to the organic carbon content of natural soils.

Disagreement justification:

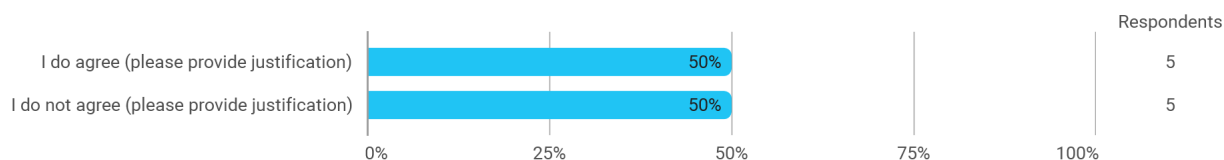
- ▶ The guideline should provide the option to use 5% peat as an alternative to 10% peat, however the 10% option should not be removed. Care should be taken that any changes to the test guideline do not invalidate previous studies, necessitating repeating significant numbers of animal studies
- ▶ For highly lipophilic substances logP correction compensates for high organic matter content. Therefore, a change of the guideline is considered as not necessary. For species other than Eisenia even higher amounts of peat may be necessary (justification needed).

- The system was calibrated/validated with 10% peat. Furthermore the use of 5% peat is already included in the guideline as optional (see note 2 of Par 11) . The current wording provides highest flexibility. Hence there is no need to change it

12. para11 - The amount of organic matter used should be decreased from 10 % to 5 %.

Rational: 10 % organic matter is not representative for many natural soils.

Furthermore, a high amount of organic matter likely reduces the bioavailability of the test chemical and therefore reflects a best case rather than a worst case scenario.



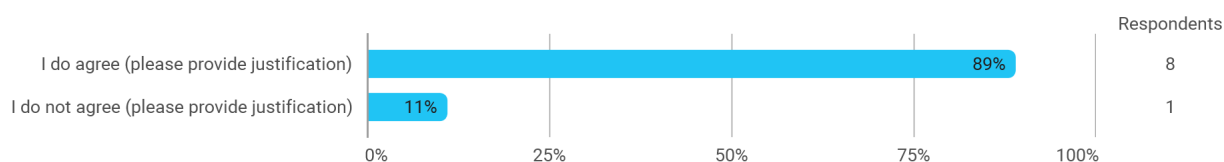
Agreement justification:

- For the reproduction test the soil similar to folsomia and hypoaspis should be used
- in accordance with suggestion 11
- We support this suggestion as 5 % organic carbon content is closer to the organic carbon content of natural soils.

Disagreement justification:

- Why? Why not keeping both options.
- The option to conduct the test with 10% peat should remain, however an additional option of 5% peat should be included.
- For highly lipophilic substances logP correction compensates for high organic matter content. Therefore, a change of the guideline is considered as not necessary. For species other than Eisenia even higher amounts of peat may be necessary (justification needed).
- The system was calibrated/validated with 10% peat. Furthermore the use of 5% peat is already included in the guideline as optional (see note 2 of Par 11) . The current wording provides highest flexibility. Hence there is no need to change it

13. para 12 - The following sentence "(corresponding to 50 +/- 10% moisture dry mass)" should be deleted



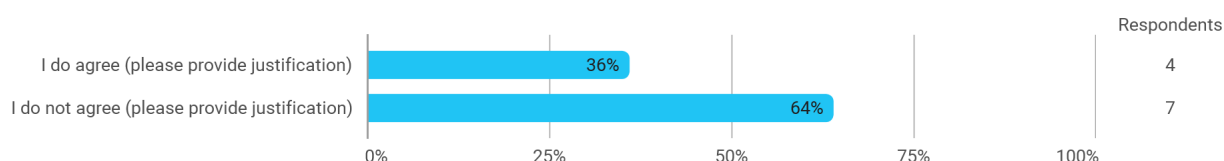
Agreement justification:

- Should be harmonised with the folsomia and hyposapis TG
- If not needed
- This sentence is unnecessary as the % MWHC values are given.
- incorrect

- The moisture dry mass may vary depending on the peat source.
- The range of the soil moisture is well described
- Agreed, as it is repetitive information.

14. para 16 - The following sentence should be changed to read as below: „During this period the worms should not be fed. “

Rational: In the past, feeding in the acclimatization period – which is normally 1 day – prior to test start has been conducted inconsistently. We suggest to limit the acclimatization period to 1 day (which is the ‘usual’ time for this period and to dismiss feeding of worms in this time frame due to different approaches of labs. If no agreement between stakeholders can be reached towards limiting the acclimatization period to 1 day and to dismiss feeding in the acclimatization period, a specification should be included in this paragraph. The information about quantity as well as quality of given food should be reported in test protocols, even if no food has been given.



Agreement justification:

- increase of the exposure at test start

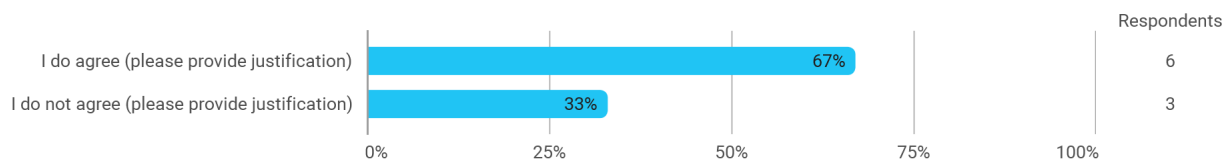
Disagreement justification:

- Acclimatisation to the test conditions should include the food provided
- Worms should be fed! Starved worms would falsify the result.
- This could add additional, unnecessary stress to the test organisms. The acclimation period is stated as “at least” one day, meaning it can be longer. It is quite common to acclimatise over a weekend prior to initiating a study on a Monday. This leads to the earthworms being without food for up to 4 days (as feeding is not conducted on Day 0). This is an unnecessarily long time without food, which causes undue suffering to the test organism. Additionally, the acclimation period is not just intended to acclimatise the earthworms to the test soil, but also to the test food, which may be different to the food used in the culture. This is an important step in the acclimation process.
- A starving period would create an unrealistic worst-case scenario as worms would consume unusually high amounts of food right after the start of the test. Also, a starvation period could cause higher control mortality effects.
- Not plausible to me why the animals should be put under starvation stress before the initiation of the study. The adaptation phase should reflect the test conditions.

15. para 30 - The water content should not vary by more than 10 % from that at the start of the test. Nevertheless, variations up to 15 % are possible and acceptable.

Rational: Throughout the 8 weeks of the study, the soil undergoes a certain development caused by the activity of the earthworms and the additional feeding.

Particularly the dried food leads to a higher dryness of the soil that should be avoided as the earthworms are very sensible towards a dry environment. Moreover, soil properties like the maximum water holding capacity have influences on the moisture of the soil throughout the study. Therefore we would appreciate to attenuate this recommendation as given above.



Agreement justification:

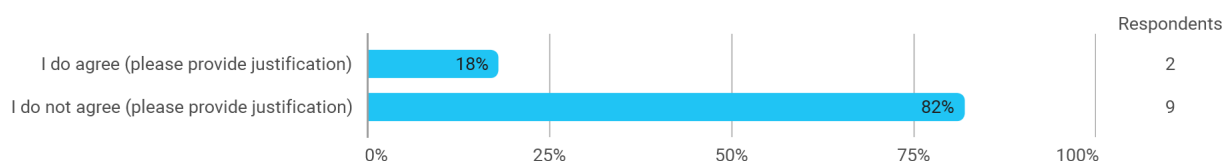
- An important point. How can we interpret this? One example: water content of 20%. What is the allowed variation if the TG states 10%: 10% - 30% or 18% - 22%?
- This allows a bit more flexibility and has no negative impact on the test
- I do agree as long it is according to Guideline limits. Anaerobic conditions due to very high moisture should be strictly avoided!
- this variation will not influence the performance of the study

Disagreement justification:

- It is not hard to achieve a moisture deviation of <10% of the starting value
- We do not agree. It is the nature of recommendations within the guideline, that deviations cannot – per se – lead to non-acceptance of the study, in contrast to the validity criteria. Extending the acceptable variation in water content will therefore only lead to a less careful monitoring of the water content. However, if within a study the water content varies up to 15 %, the study can still be acceptable upon judgement of the risk assessor.

16. para 33 - It should be added that the food spread on the soil surface must be treated with the tested substance at the corresponding concentrations.

Rational: The suggestion follows the “Scientific Opinion addressing the state of the science on risk assessment of plant protection products for in-soil organisms”, EFSA (2017), according to which “exposure via the litter layer is [...] a relevant route of exposure for some soil invertebrates, like macroarthropods, slugs and snails. [...] This may lead to underestimation of internal exposure due to dilution.” An underestimation of effects has been shown by Natal-da-Luz et al. (2019)¹ in a comparison of sensitivity of *Hypoaspis aculeifer* when being fed with contaminated or non-contaminated food in addition to exposure via soil. *H. aculeifer* showed higher sensitivity to the tested substance when fed with contaminated food (mites). These results suggest that oral exposure should be taken into account and could be transferred to the earthworm test system. 1 Natal-da-Luz, T., Gevaert, T., Pereira, C., Alves, D., Arena, M., & Sousa, J. P. (2019). Should oral exposure in *Hypoaspis aculeifer* tests be considered in order to keep them in Tier I test battery for ecological risk assessment of PPPs?. *Environmental pollution*, 244, 871-876.



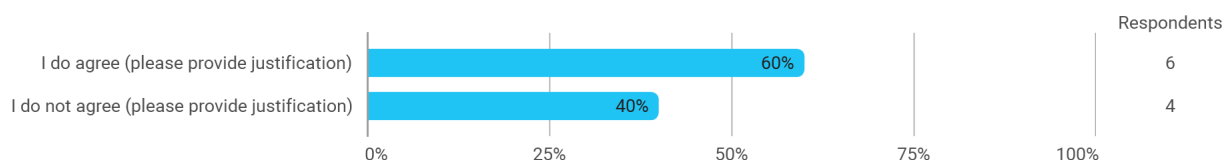
Agreement justification:

- Totally agree. See comment on point 14.

Disagreement justification:

- Under field conditions, the worms also would have uncontaminated food available
- Definitely not useful!!
- clear procedure should be described
- Exposure via contact and exposure via ingestion are two different pathways. It would be more appropriate to consider exposure during feeding in a separate study, so that toxicity can be attributed to the appropriate exposure pathway and thereby can be contextualised within the risk assessment. It should be noted that significant research needs to be undertaken prior to consideration of the need for an oral exposure test for earthworms, to understand the relative importance of oral versus dermal exposure pathways for earthworms.
- This is not a feeding test.
- Treated food would represent an unrealistic worst-case scenario. Under field conditions, soil organic matter does not necessarily contain the same concentrations as the surrounding soils. High amounts of uncontaminated organic matter reach the soil in the process of incorporation of crops after harvest.
- clear procedure should be described
- The test system was validated as is. If different exposure scenarios should be included the technical feasibility and reproducibility needs to be checked again. Furthermore accurate application of the food item will be challenging and need to be evaluated in order to gain an understanding on the accuracy/variability and therefore reproducibility

17. para 33 - After the first sentence an additional sentence should be added as in the following: "If the test substance is mixed into the soil the first food ration can also be provided on the day of application"



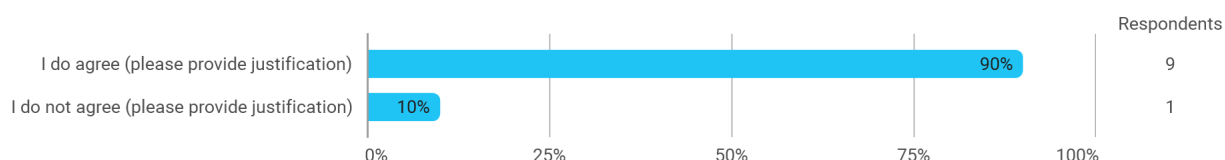
Agreement justification:

- the interval of feeding would be more evenly
- already standard
- I do agree, flexibility for food provision is okay.

Disagreement justification:

- ▶ The intention of delayed feeding is to encourage the earthworms to move through the test soil and thus ensure exposure to the test substance. Providing food immediately could affect the outcome of the study.
- ▶ Food should be added to the soil one day before test start and one day after test start.
- ▶ I do not see the need to change this

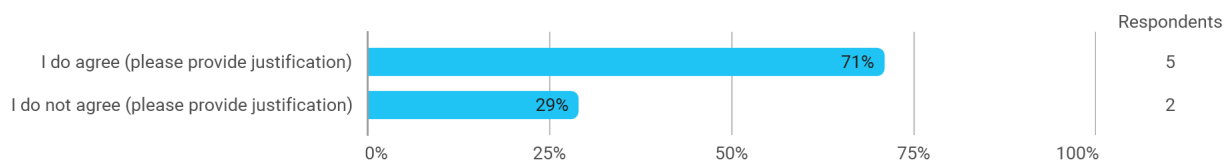
18. para 36 - The following sentence must be clarified: "For determination of the NOEC, at least five/twelve concentrations in a geometric series should be tested." It is not clear whether five or twelve concentrations should be tested.



Agreement justification:

- ▶ What is the requirement to use 5 or 12 doses?
- ▶ more than 5 test concentrations are already standard
- ▶ Clarification would be beneficial
- ▶ Twelve must be deleted
- ▶ I agree that it is not specified properly. However, guideline flexibility is needed regarding choice of concentrations and number of concentrations in order to design properly for each substance depending on the dose-response behaviour. Even a limit test should be possible, if needed and suitable for non-toxic test substances.
- ▶ clarification is helpful

19. para 36 - First paragraph: the word "twelve" should be deleted



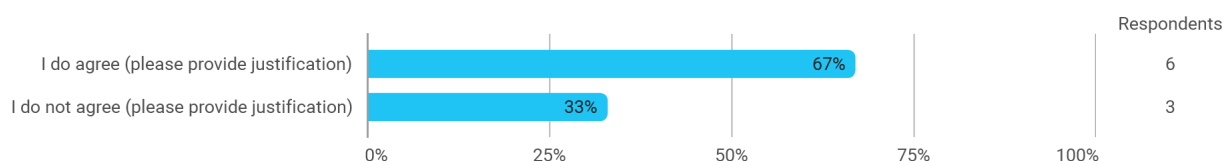
Agreement justification:

- ▶ 5 concentrations are sufficient to determine a NOEC
- ▶ Clarification would be beneficial
- ▶ Twelve must be deleted

Disagreement justification:

- ▶ I do not agree. There should be no minimum in number of concentrations.

20. para 39 - In the first sentence the words " and cocoon numbers" should be deleted



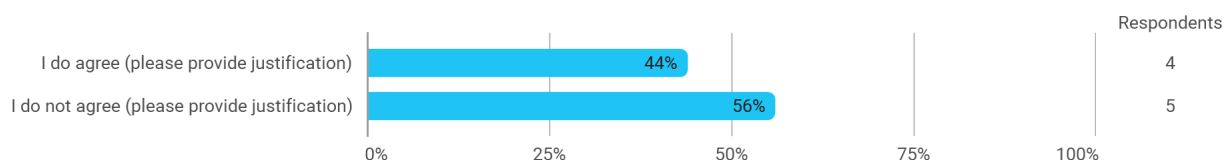
Agreement justification:

- The parameter reproduction is best described by the number of juveniles
- information not needed
- Or should be at least mentioned as optional

Disagreement justification:

- Instead, this wording needs to be changed to “number of unhatched cocoons”.
- We do not agree. Instead, as cocoon number are normally counted in test laboratories, the number of cocoons should be used as an additional reproductive endpoint.

21. para 39 - The last sentence of paragraph 39 should be moved on the end of paragraph 37



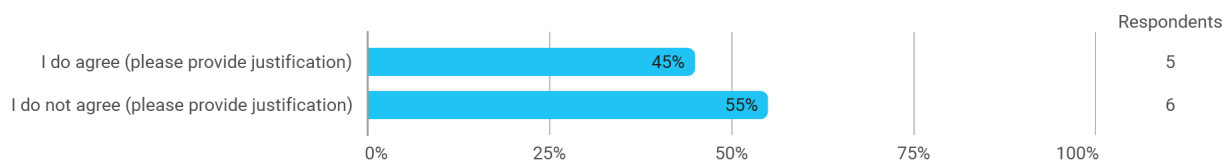
Agreement justification:

- Even better, it should be an own para at the end of this chapter
- it is already mentioned there
- Agreement. In fact, the last sentence of paragraph 39 should replace the last sentence of paragraph 37.

Disagreement justification:

- Para 37 describes the handling of adult worms
- It is good where it is.
- Paragraph 37 deals with the necessary steps on Day 28, whereas the last sentence of paragraph 39 is in relation to the necessary steps on the last day of the in-life phase.
- Not possible, as no juveniles are present at day 28.

22. para 42 - Should be deleted at all.



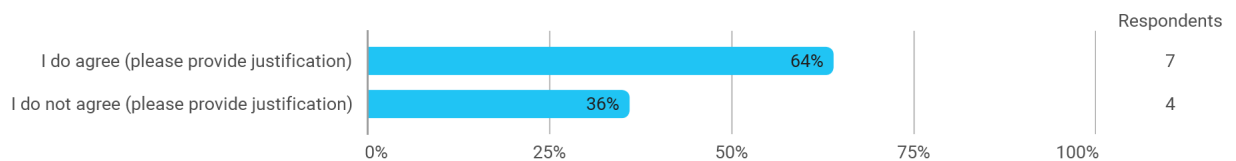
Agreement justification:

- Mortality should not occur in e reproduction test
- The endpoint mortality is an integral part and included in the asset of the number of juveniles. It is therefore a good proposal to streamline the evaluation and delete this part.

Disagreement justification:

- Mortality is a relevant endpoint, therefore the treatment of the results should be remained
- See point 2 above. The primary intention of OECD 222 is to assess sub-lethal effects. However, if mortality occurs it should be documented and assessed. Paragraph 42 provides guidance on this.
- No. The first two sentences of the paragraph should be kept and only the following statistical suggestions can be deleted.
- We do not agree. As mortality can be a more sensitive endpoint in certain cases/for certain chemicals than reproduction, this endpoint must be kept and used for risk assessment.

23. para 42 - Evaluation of biomass should be added.



Agreement justification:

- Weight change is important
- Similar to mortality, biomass can provide useful insight into the study outcome, therefore guidance on how to present and interpret these data based on scientific understanding of the population relevance of biomass would be useful. However, it should be reiterated that Biomass is not a relevant endpoint of the study, merely useful for context.
- evaluation of biomass is missing
- Totally agree. We originally made comment number 25, which is exactly the same.

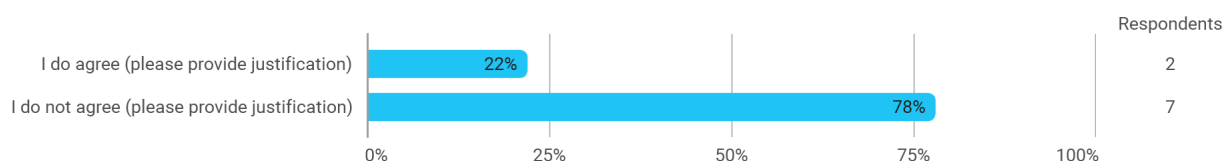
Disagreement justification:

- Weight change is a relevant parameter and should have a own paragraph
- ECx evaluation for biomass and growth do not make sense! Growth alternates between positive and negative values. Biomass values are biologically limited (max body size) and biomass changes alternate between positive and negative values as well.
- In comparison to mortality/number of juveniles the endpoint biomass is often less reproducible. Reproduction data are more reliable parameter

24. para 42 - In a new paragraph, the endpoint “number of cocoons” should be included as a relative endpoint, based on the number of cocoons per replicate/juvenile worm after four weeks.

Rational: This additional parameter would add a sublethal endpoint that could for certain substances be more sensitive than the fecundity endpoint based on number of

juveniles. As the number of cocoons is already required to be counted according to OECD 222 (2016), paragraph 39, this would not require additional laboratory working steps but would lead to the detection of a broader range of population-relevant effects.



Agreement justification:

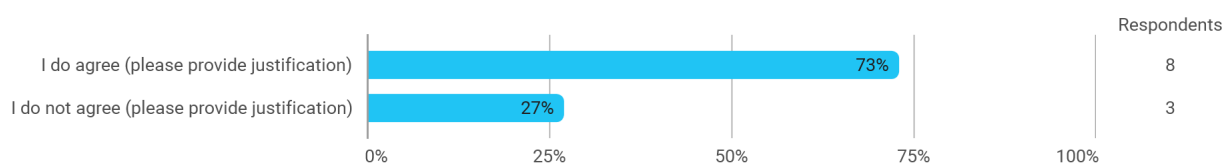
- ▶ Totally agree. We originally made this comment.

Disagreement justification:

- ▶ Number of cocoons does not reflect the reproductive output sufficiently
- ▶ Definitely not, cocoons as endpoint are meaningless
- ▶ See point 20 above. The number of unhatched cocoons should be reported. It is not clear how useful it would be to assess all cocoons when juveniles are already being assessed. This is double counting.
- ▶ no additional information
- ▶ The number of cocoons is not necessarily an indicator for reproduction since the number juveniles can differ per cocoon. Number of juveniles is the better measurement of reproductive output as juveniles might die after hatching.
- ▶ the number of juveniles provide a robust endpoint for reproduction data.

25. para 42 - In a new paragraph, the endpoint “biomass” should be included as a relative endpoint, based on the weight change of adult worms after four weeks.

Rational: This additional parameter would add a sublethal endpoint that could be more sensitive for certain substances than the endpoint mortality, also measured after four weeks. As the changes in body weight are already required to be calculated according to OECD 222 (2016), paragraph 50, this would not require additional laboratory working steps but would lead to the detection of a broader range of population-relevant effects.



Agreement justification:

- ▶ Weight change is a relevant parameter and should have a own paragraph
- ▶ Weight change is important
- ▶ already standard

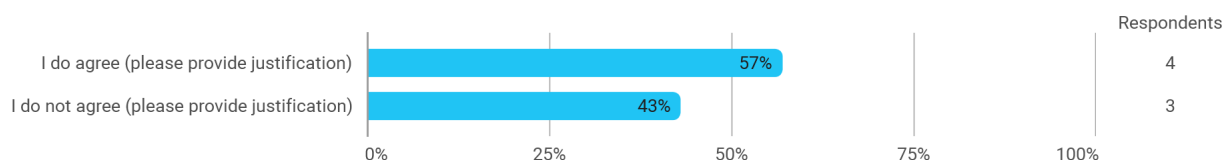
- See point 23. Guidance on how to present biomass data would be helpful, however it should not be an endpoint from the study as it is unclear what the population relevance of this endpoint is.
- see suggestion 23

Disagreement justification:

- It is scientifically not appropriate to derive an ECx curve or value from a relative endpoint.
- In comparison to mortality/number of juveniles biomass is often less reproducible. Reproduction data are more reliable and in general the more sensitive parameter

26. Annex 5: Requirements regarding the efficiency of the extraction of the predatory mites should be updated according to ISO 21285 and the publication of Huguier P., Manier N., Pandard P. (2016) Evaluation of the extraction efficiency for the Hypoaspis aculeifer reproduction test in the context of soil quality assessment. Ecotoxicology 25: 1867–1872.

Rational: It is indeed important to consider adults and juveniles separately with regard to the extraction efficiency. The following requirement has been adopted in ISO 21285: "The extraction efficiency of the chosen method shall have been proven once or twice a year using vessels containing a known number of juvenile and adult mites kept in untreated test substrate. Efficiency should be $\geq 90\%$ on average for adult mites while this value is lowered to 75 % for juveniles.



Agreement justification:

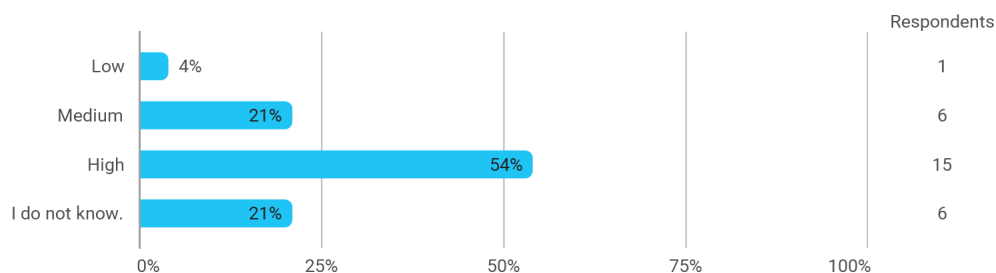
- The extraction method should be completed as it's very vague
- I do agree. However, further guidance and validation is needed. A ring test is needed to define threshold values for acceptable extraction values for earthworms.

Disagreement justification:

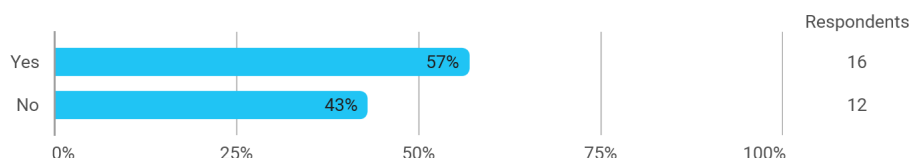
- This is not relevant to the discussion of OECD222.
- wrong guideline

B.17 Test No. 226: Predatory mite (*Hypoaspis* (*Geolaelaps*) *aculeifer*) reproduction test in soil

Relevance of revision



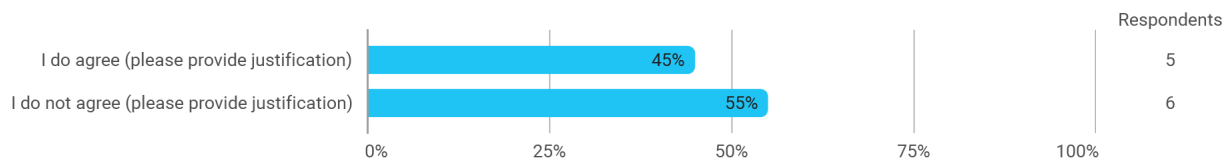
Interested to be part of an OECD Expert Group



The following suggestions have been collected:

1. "Determination of test substance concentration at appropriate time intervals (at least beginning and end of test)" should be included.

Rational: Calculated EC-values will be based on actual concentration instead of the nominal concentration. Such an approach is established for aquatic test systems and it is highly recommended to use a similar approach for soil organisms.



Agreement justification:

- A dose verification at the beginning of the test is desirable
- To confirm the test concentrations.
- Total agreement. We originally commented on point 7 which suggests basically the same.

Disagreement justification:

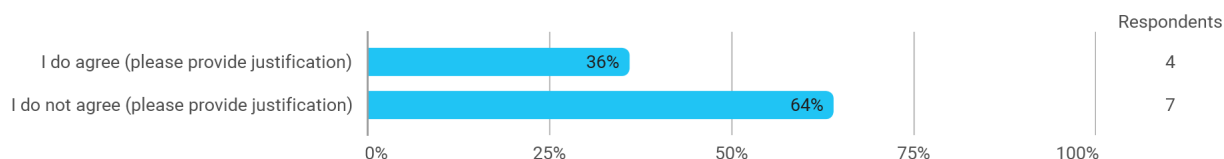
- It needs to be defined and explained how and why this should be necessary
- The appropriate guideline for understanding the fate and behaviour of the test substance in soil is OECD 307. Additionally, the use of artificial soil in the hypoaspis reproduction study would mean such information has little relevance to real world scenarios. A more appropriate approach is to analytically confirm dosing e.g. by analysis of the dosing solutions. The determination of a mean measured soil concentration or a twa is not appropriate for this study type. Not least because it is not possible to design a study type that allows concentrations to be maintained e.g. a flow through system. The risk assessment is conducted by comparing the peak initial concentration from the exposure models with the

effects that are seen at that peak concentration. The current test design reflects the risk assessment requirements.

- This is a topic which must be clarified for all in-soil organism test guidelines in general, not in individual guidelines
- In case studies are performed under GLP conditions and Certificates of Analysis are provided for each test substance, the amounts of substances or products applied do not need to be analytically verified. Moreover, analytical measurements in GLP soil laboratory tests are in many cases not meaningful due to multiple analytical difficulties when extracting from the soil matrix.
- Analytics will increase costs with limited benefit. There might be specific test thesis where analytic is helpful. Historically the old approach is sufficient and provides reliable information. Before an inclusion of analytics the use of the information should be discussed also in the context of a potential risk assessment. It should be kept in mind that the compulsory inclusion would lead to a huge repetition of studies unless there are considerations that old studies with no analytics are still valid

2. It might be considered to make the requirement for analytical check for the behaviour of the chemical compulsory. The comment might be relevant for all tests using artificial soil.

Rational: It is mandatory for aquatic organisms (even for acute tests), including sediment dwellers. Indeed, the dissipation of the chemical largely influences the effect, therefore the endpoint get from the test (which is the main aim). The dissipation rate can considerable different in the test media than in the field. Therefore, the use of the endpoints to compare the toxicity of the different chemicals might be a questionable practice. What the endpoint actually express is a combination of the toxicity and the dissipation rate of the chemical in the test media



Agreement justification:

- To confirm the test concentrations.
- Disagreement. We understand this comment as a suggestion to better check the properties “volatile, unstable or readily degrading substances” in order to judge whether analytical monitoring must be performed. This would in general be a good suggestion, however, we suggest performing analytical monitoring for all substances.

Disagreement justification:

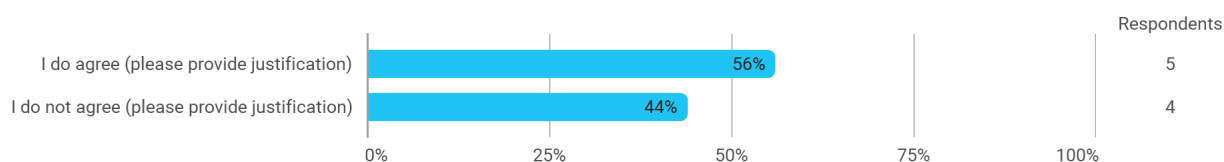
- The behaviour of the test item can be assessed with the TG 307
- It needs to be defined and explained how and why this should be necessary
- The appropriate guideline for understanding the fate and behaviour of the test substance in soil is OECD 307. Additionally, the use of artificial soil in the hypoaspis reproduction study would mean such information has little relevance to real world scenarios. A more

appropriate approach is to analytically confirm dosing e.g. by analysis of the dosing solutions. The determination of a mean measured soil concentration or a twa is not appropriate for this study type. Not least because it is not possible to design a study type that allows concentrations to be maintained e.g. a flow through system. The risk assessment is conducted by comparing the peak initial concentration from the exposure models with the effects that are seen at that peak concentration. The current test design reflects the risk assessment requirements.

- ▶ See comment on suggestion 1
- ▶ In case studies are performed under GLP conditions and Certificates of Analysis are provided for each test substance, the amounts of substances or products applied do not need to be analytically verified. Moreover, analytical measurements in GLP soil laboratory tests are in many cases not meaningful due to multiple analytical difficulties when extracting from the soil matrix.
- ▶ Analytics will increase costs with limited benefit. There might be specific test thesis where analytic is helpful. Historically the old approach is sufficient and provides reliable information. Before an inclusion of analytics the use of the information should be discussed also in the context of a potential risk assessment. It should be kept in mind that the compulsory inclusion would lead to a huge repetition of studies unless there are considerations that old studies with no analytics are still valid

3. **An alternative, optional test design should be included to allow data generation for parameterisation of toxicokinetic-toxicodynamic models (e.g. GUTS and DEB-TKTD). This alternative design should include additional measurements of mortality and reproductive output at intermediate time points (at least 2). It should also include measurements of organismal size at four time points (day 0, intermediate x 2, day 14). This can be achieved by destructive sampling of test vessels. Replication at each sampling time is not necessary for parameterisation of toxicokinetic-toxicodynamic models. Instead replicate test vessels can be re-distributed for sampling at intermediate time points.**

Rational: Toxicokinetic-toxicodynamic modelling (GUTS and DEB-TKTD), especially when combined with population modelling, can help to quantify and reduce uncertainties in environmental risk assessment by providing improved integration of exposure and effects and by incorporating different pesticide application scenarios: EFSA Panel on Plant Protection Products their Residues, C. Ockleford, P. Adriaanse, P. Berny, T. Brock, S. Duquesne, S. Grilli, A. F. Hernandez-Jerez, S. H. Bennekou, and M. Klein. 2018. Scientific Opinion on the state of the art of Toxicokinetic/Toxicodynamic (TKTD) effect models for regulatory risk assessment of pesticides for aquatic organisms. 1831-4732.; SCENIHR (Scientific Committee on Emerging and Newly Identified Health Risks), SCHER (Scientific Committee on Health and Environmental Risks), SCCS (Scientific Committee on Consumer Safety). 2012. Addressing the new challenges for risk assessment. Brussels: European Commission.



Agreement justification:

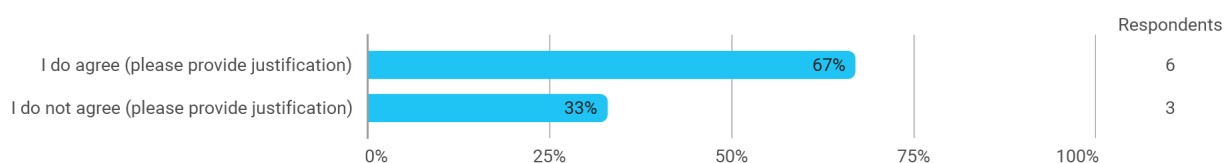
- ▶ Modelling can assist to provide further data, where experimental phases in the lab are limited
- ▶ The suggested updates to these guidelines will enable more realistic and ecologically relevant environmental risk assessments.
- ▶ However further detailed considerations need to be done to define the needed parameters. In order to not overcomplicate the guideline it might be an option to include the alternative design in an appendix.

Disagreement justification:

- ▶ This is a reproduction test
- ▶ I do not agree to include this as a mandatory test design in the main chapters. Modelling is not an useful tool for all substances due to low toxicity of many substances. It must be included as Appendix (optional test design).
- ▶ Due to high variability and uncertainty associated with the 3-dimensional patterns of toxicant degradation and organism movement in soil (“patchiness”), exposure profiles for soil organisms can be predicted only with high uncertainty. This is already a key issue for Tier 1 experimental testing. Therefore, the use of TKTD models to predict effects on soil organisms in time and space as in the aquatic Tier 2C - where organism exposure in the water column or at the sediment surface is potentially more homogeneous in space- is highly questionable. Developing reliable TKTD models for predicting effects on soil organisms and setting up meaningful scenarios for TKTD modelling in order to identify a realistic worst-case would require huge (even unrealistic) efforts in characterizing the toxicant degradation and the test organism’s behaviour; the species tested should also be suitable to represent of potential toxicologically sensitive and ecologically vulnerable species under realistic field conditions. However, TKTD models could be more successfully applied to retrospectively analyse and understand differences in effects on soil organisms observed across different studies, e. g. across soils having different characteristics. In general, we agree that analytical measurements should be performed in laboratory studies in order to better link exposure to effects. In this regard, residues should be analyzed at different exposure levels in the soil, soil pore water and organisms at the beginning, during and at the end of the study.

4. para 3 - Suggestion to bring in line with OECD 232, where it is more explicit in this section that mortality is also a relevant endpoint of the study.

Rational: For OECD 222 and 226 it sometimes is a point of discussion whether mortality should be determined as endpoint for the study or not. In addition, there is a discrepancy between OECD 222 + 226 on the one hand, and OECD 232 on the other hand, as in OECD 232 it seems clear that mortality should be used as well for endpoint derivation, whilst in OECD 222 and 226 this is not clear.



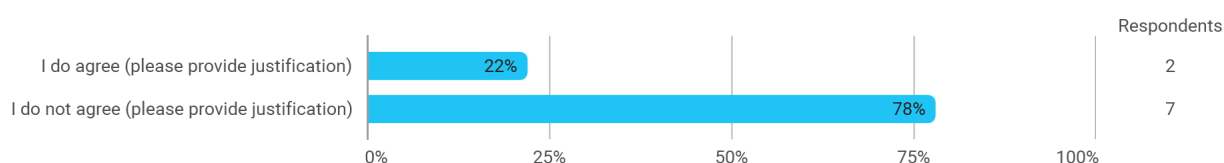
Agreement justification:

- ▶ TG of the same ecological compartment should be harmonised
- ▶ Mortality should continue to be assessed and reported, however changes to paragraph 3 to make mortality a required endpoint would necessitate changes throughout the guideline, which should be undertaken with caution. The focus of this study should remain on sub-chronic effects, particularly reproduction. If the test guideline is amended to make mortality a required endpoint, there should also be an acknowledgement that the maximum application rate tested should be focused around the expected sub-chronic toxicity range and not the mortality range, i.e. mortality as an additional useful parameter, but the primary emphasis of this study design should continue to be to detect effects on reproduction.
- ▶ I do agree.
- ▶ Agreement. For certain chemicals, mortality can be a more sensitive endpoint in certain cases than reproduction. Therefore, the endpoint for mortality should always be calculated.

Disagreement justification:

- ▶ I do not agree
- ▶ Mortality should not occur in reproduction tests
- ▶ Reproduction data (number of juvenile) will include lethal mortality aspects (only survived animals will reproduce). As reproductive data cover the mortality and is the more sensitive endpoint as well there is no need to focus on mortality endpoints

5. **para 5 - This is a more general question for soil ecotoxicology studies: the data requirements for active substances section 4 (analytical methods) states that: "Methods shall be submitted, with a full description, for the determination of non-isotope-labelled residues in all areas of the dossier, as set out in detail in the following points: in soil, water, sediment, feed and any additional matrices used in support of ecotoxicology studies;" However in practice, analytical methods for soil ecotoxicology studies are almost never provided, since it is not a direct requirement in the guideline. it is still defined as "optional": For volatile, unstable or readily degrading substances (e.g., data generated from a TG 307 study may be considered), or where there is otherwise uncertainty in maintaining the nominal soil concentration, analytical measurements of the exposure concentrations at the beginning, during and at the end of the test should be considered. Would it be possible to clarify that analytical verification of test concentrations in soil is required in the guideline?**

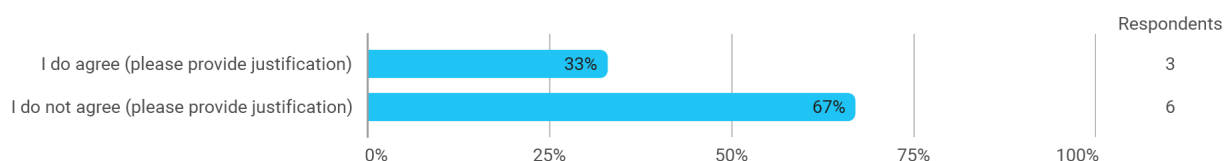


Agreement justification:

- ▶ Total agreement. We made a similar suggestion with point 7.

Disagreement justification:

- ▶ Clarify first (analytical dose verification, soil extraction, biodegradation...)
 - ▶ analytical verification is not required
 - ▶ The appropriate guideline for understanding the fate and behaviour of the test substance in soil is OECD 307. Additionally, the use of artificial soil in the Hypoaspis reproduction study would mean such information has little relevance to real world scenarios. A more appropriate approach is to analytically confirm dosing e.g. by analysis of the dosing solutions.
 - ▶ General topic, not for individual guidelines
 - ▶ In case studies are performed under GLP conditions and Certificates of Analysis are provided for each test substance, the amounts of substances or products applied do not need to be analytically verified. Moreover, analytical measurements in GLP soil laboratory tests are in many cases not meaningful due to multiple analytical difficulties when extracting from the soil matrix.
 - ▶ An inclusion of analytics should be justified by reasonable means. Studies were performed without analytics without any significant issues. Analytics would again increase the costs and complexity. There might be specific questions where analytics is justified, however it should remain as an optional point
6. **para 5 - Further guidance on follow up of the last sentence of this paragraph ('For volatile (...) considered.') is considered necessary. I.e. : criteria to be used for deciding when a substance is considered 'unstable' (e.g. a DT50-value) when measurements are performed, criteria for when to use nominal or measured concentrations (probably the 80% criterium) how to express measured concentrations (i.e. geomean, twa) when no measurements are performed while considered necessary based on the above, what are the consequences for the validity of the study.**
- Rational: The 'strictness' of the last sentence of this paragraph ('For volatile (...) considered.'), is not very clear and it leaves open several questions.**



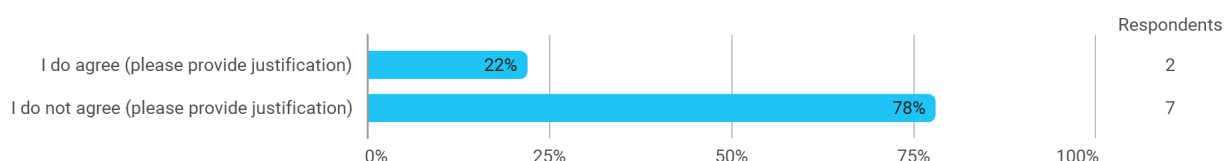
Agreement justification:

- ▶ Further guidance how to handle volatile substances is needed

Disagreement justification:

- ▶ The determination of a mean measured soil concentration or a twa is not appropriate for this study type. Not least because it is not possible to design a study type that allows concentrations to be maintained e.g. a flow through system. The risk assessment is conducted by comparing the peak initial concentration from the exposure models with the effects that are seen at that peak concentration. The current test design reflects the risk assessment requirements.
- ▶ General topic, not for individual guidelines

- ▶ In case studies are performed under GLP conditions and Certificates of Analysis are provided for each test substance, the amounts of substances or products applied do not need to be analytically verified. Moreover, analytical measurements in GLP soil laboratory tests are in many cases not meaningful due to multiple analytical difficulties when extracting from the soil matrix.
 - ▶ It should be kept in mind that detection of residues in soil has different complexity than for example in the aquatics where the 80% criteria is used. In general analytics in soil lab studies will lead to a lot of challenges regarding the interpretation.
 - ▶ Partly agree. However, we suggest performing analytical monitoring for all substances.
7. **para 5 - The following sentence should be deleted: “For volatile, unstable or readily degrading substances (e.g. data generated from a TG 307 study may be considered), or where there is otherwise uncertainty in maintaining the nominal soil concentration, analytical measurements of the exposure concentrations at the beginning, during and at the end of the test should be considered.” Instead, analytical measurements should be performed for all substances, covering the concentration range. Here, the concentration in total soil as well as the concentration in pore water should be determined. Following the EFSA Scientific Opinion on in-soil organisms (2017), “it is recommended that the concentrations of active substances in soils is measured more than twice (see also the EFSA DegT50 EFSA Guidance), depending on the degradation of the active substances and the length of the test.” Suggested new phrasing: “For all substances, analytical measurements of the exposure concentrations at the beginning, during and at the end of the test should be considered. If the concentration of the substance being tested has been satisfactorily maintained within $\pm 20\%$ of the nominal or measured initial concentration throughout the test, analysis of the results can be based on nominal or measured initial values. If the deviation from the nominal or measured initial concentration is not within the range of $\pm 20\%$, analysis of the results should be based on geometric mean concentration during exposure or on models describing the decline of the concentration of the test substance.”**
- Rational: According to EFSA Scientific Opinion on in-soil organisms (2017, chapter 7.11.3), analytical measurements for tests on in-soil organisms are requested. In this Scientific Opinion, it is stated that “During laboratory handling procedures of the spiked soils, however, possible losses of the pesticides due to volatilization, degradation, and sorption to, e.g. the glass matrix of vessels used, may occur.” Therefore, the actual concentration of the tested substance needs to be determined in order to correctly link exposure and effects. Moreover, analytical measurements in test systems are included in all aquatic studies. Terrestrial studies should be analyzed in the same way, otherwise ‘real’ exposure cannot be described and an extrapolation from lab to field is questionable.**



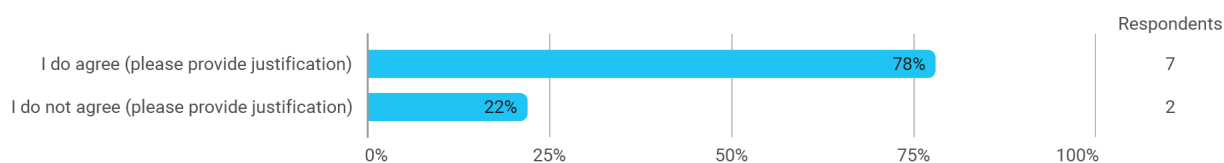
Disagreement justification:

- ▶ What are the consequences for the validity of the study. How to handle the control, when geometric mean must be used.
- ▶ Why should this be necessary?
- ▶ The appropriate guideline for understanding the fate and behaviour of the test substance in soil is OECD 307. Additionally, the use of artificial soil in the hypoaspis reproduction study would mean such information has little relevance to real world scenarios. A more appropriate approach is to analytically confirm dosing e.g. by analysis of the stock solutions.
- ▶ General topic, not for individual guidelines
- ▶ In case studies are performed under GLP conditions and Certificates of Analysis are provided for each test substance, the amounts of substances or products applied do not need to be analytically verified. Moreover, analytical measurements in GLP soil laboratory tests are in many cases not meaningful due to multiple analytical difficulties when extracting from the soil matrix.
- ▶ An inclusion of analytics should be justified by reasonable means. Studies were performed without analytics without any significant issues. Analytics would again increase the costs and complexity and would only provide limited information. There might be specific questions where analytics is justified, however it should remain as an optional point. It should be kept in mind that detection of residues in soil has different complexity than for example in the aquatics where the 80% criteria is used. In general analytics in soil lab studies will lead to a lot of challenges regarding the interpretation.

8. para 10 - Alternatives to sphagnum peat as a component of OECD artificial soil should be investigated

Rational: Renewable alternatives to sphagnum peat as a component of OECD artificial soil should be investigated given the environmental concerns associated with peat harvesting. De Silva & Van Gestel (2009) suggest composted coco peat as a suitable replacement. In horticulture, sphagnum peat in potting compost is increasingly being replaced with alternatives such as coir (i.e. coco peat) - how much longer will sphagnum peat be readily available for?

De Silva, P.M.C.S., Van Gestel, C.A.M., 2009. Development of an alternative artificial soil for earthworm toxicity testing in tropical countries. Appl. Soil Ecol. 43, 170–174.



Agreement justification:

- ▶ Peat moss is not sustainable
- ▶ In the interests of sustainability alternatives to peat should be actively investigated.
- ▶ sphagnum peat very variable

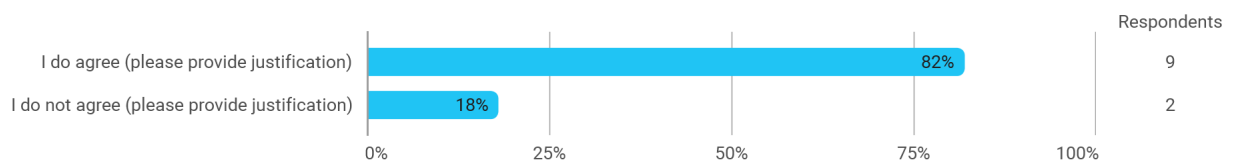
- The extraction of peat contributes directly to climate change. In contrast, using alternatives would contribute to species conservation and climate protection. In the future, the extraction of peat could also be prohibited by law for climate protection reasons. Therefore, it is to be supported to look for alternatives soon.

Disagreement justification:

- Since the artificial OECD soil provided reliable and reproducible results, I do not see the big need for alternatives. Regardless a search for new alternative might be appropriate but would also need for additional feasibility/ring testing.

9. para 10 - An alternative to sphagnum peat as source of organic matter should be provided.

Rational: The OECD artificial soil requires 5 % organic matter. It is recommended to use sphagnum peat. However, sphagnum peat is not available worldwide and its use is also questionable for ecological reasons.



Agreement justification:

- Peat moss is not sustainable
- limited availability of peat
- In the interests of sustainability alternatives to peat should be actively investigated
- General topic, not for individual guidelines
- I agree. However, an alternative to sphagnum peat as source of organic matter can only be provided after a feasible validation and ring-test process.
- limited availability of peat
- The extraction of peat contributes directly to climate change. In contrast, using alternatives would contribute to species conservation and climate protection. In the future, the extraction of peat could also be prohibited by law for climate protection reasons. Therefore, it is to be supported to look for alternatives soon.

Disagreement justification:

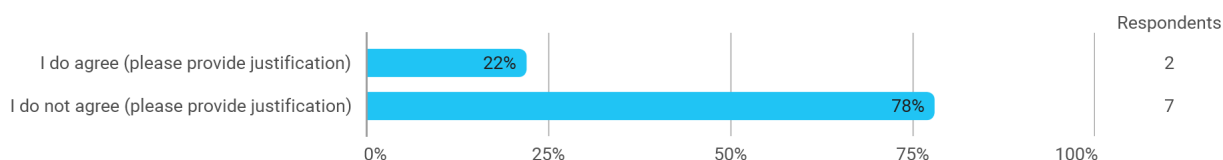
- Since the artificial OECD soil provided reliable and reproducible results, I do not see the big need for alternatives. Regardless a search for new alternative might be appropriate but would also need for additional feasibility/ring testing.

10. para 25 - It should be added that the food added to each container should be treated with the tested substance at the corresponding concentrations.

Rational: The suggestion follows the “Scientific Opinion addressing the state of the science on risk assessment of plant protection products for in-soil organisms”, EFSA (2017), according to which “exposure via the litter layer is [...] a relevant route of exposure for some soil invertebrates, like macroarthropods, slugs and snails. [...] This

may lead to underestimation of internal exposure due to dilution. An underestimation of effects has been shown by Natal-da-Luz et al. (2019)¹ in a comparison of sensitivity of *Hypoaspis aculeifer* when being fed with contaminated or non-contaminated food in addition to exposure via soil. *H. aculeifer* showed higher sensitivity to the tested substance when fed with contaminated mites. These results suggest that oral exposure should be taken into account for the *H. aculeifer* test system.

¹Natal-da-Luz, T., Gevaert, T., Pereira, C., Alves, D., Arena, M., & Sousa, J. P. (2019). Should oral exposure in *Hypoaspis aculeifer* tests be considered in order to keep them in Tier I test battery for ecological risk assessment of PPPs?. *Environmental pollution*, 244, 871-876



Agreement justification:

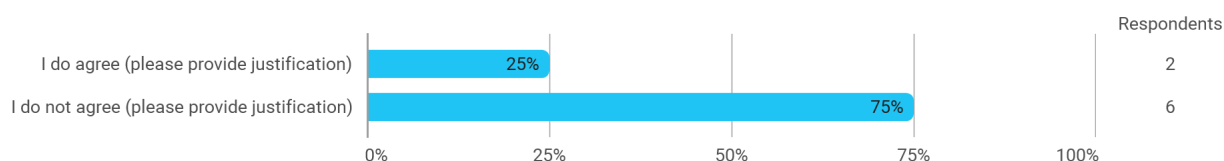
- Totally agree. We originally made this comment.

Disagreement justification:

- Under field conditions, soil mites would have uncontaminated food as well. As the food is mostly cheese mites or nematodes the impact on these organisms would have to be investigated first
- Definitely not!
- Exposure via contact and exposure via ingestion are two completely different pathways. Significant research needs to be undertaken prior to introducing a change of this sort. Firstly, to understand the major exposure pathways for *hypoaspis* and secondly to understand the impact of this change on the test output and subsequently how those data would be contextualised in a risk assessment. Finally, a change of this kind would require ring testing. It would be more appropriate to consider exposure during feeding in a separate study, so that any toxicity can be attributed to the appropriate exposure pathway and therefore contextualised within the risk assessment. Additionally, *Hypoaspis* are fed cheese mites or *Collembola*. How could the test item be mixed in with food when the food comprises live arthropods? Furthermore, treating the 'food' may lead to confounded effects if there is direct toxicity to the 'feed' organisms. Considerable research and ring testing would be required for a change of this nature.
- This would be a special test setup,
- In case provided food (e.g. nematodes or mites) will be contaminated, the food could die and would then not be eaten by the predatory test organism. In this case you would measure an indirect effect (starvation of *Hypoaspis*) and not the direct substance effect.
- The test system was validated as is. If different exposure scenarios should be included the technical feasibility and reproducibility needs to be checked again. Furthermore accurate application of the food item will be challenging and need to be evaluated in order to gain an understanding on the accuracy/variability and therefore reproducibility

11. para 32 - Change the sentence: "extraction efficiency must be validated once or twice a year in" to "extraction efficiency must be validated"

Rational: to be in line with the OECD TG232 and because there is no reason why the validated extraction method could not be reliable after a year



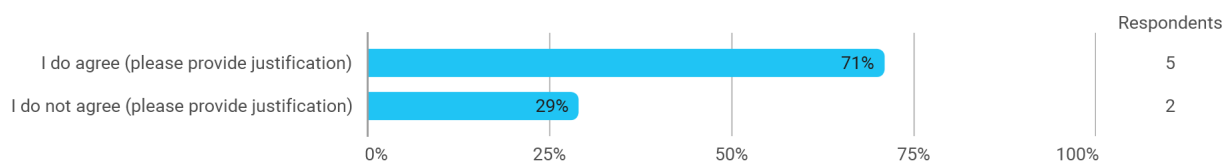
Agreement justification:

- once a year should be more than sufficient

Disagreement justification:

- extraction efficiency should be validated regularly
- Stick to once or twice a year
- It is helpful to continue to include guidance on the frequency of extraction efficiency validation.
- It is sufficient to validate extraction efficiency once or twice a year.
- Disagreement. See comment no. 26 regarding OECD 222 which has been falsely commented regarding OECD 222 even though it is a comment on OECD 226.

12. para 31 - Question: How should the behaviour be observed when the mites are in the soil during exposure and fixated in the fixation solution (and thus dead) after extraction?



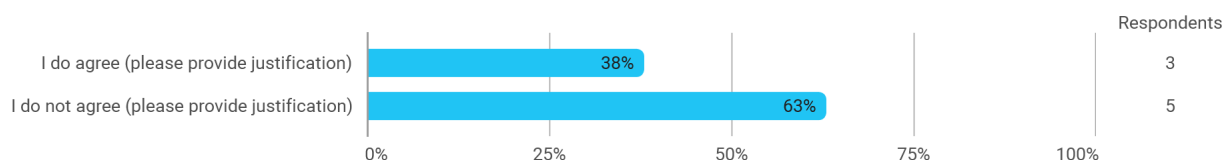
Agreement justification:

- Good point. Only changes in morphology can be reported.
- correct, not possible
- I agree that a behavioural endpoint is not relevant.

Disagreement justification:

- This does not appear to be a suggestion for the guideline revision, but rather someone asking a point for clarification

13. para 32 - Delete sentence: "At the end of the test, adults and juveniles are humanely euthanized, ..."



Agreement justification:

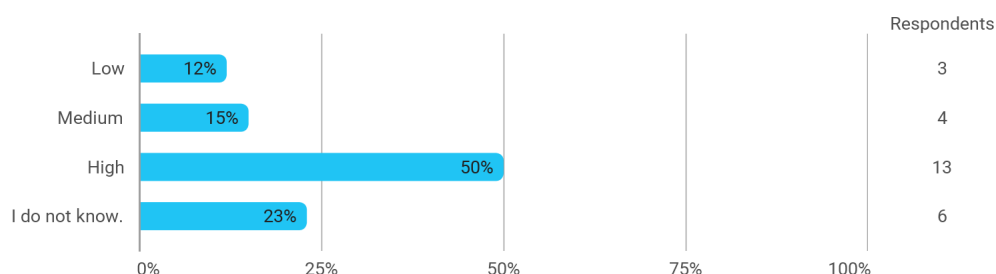
- The animals are already dead when counted in the fixation solution

Disagreement justification:

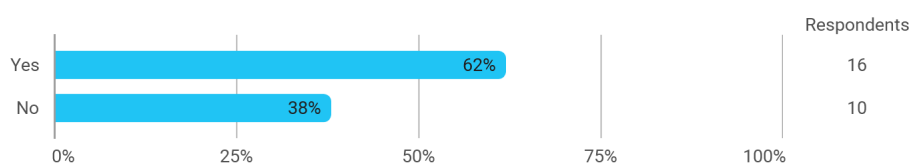
- Do not delete advise
- This seems to be a necessary sentence and is in line with the other soil organism TGs
- I do not agree. There should be a suitable procedure used.

B.18 Test No. 232: Collembolan Reproduction Test in Soil

Relevance of revision



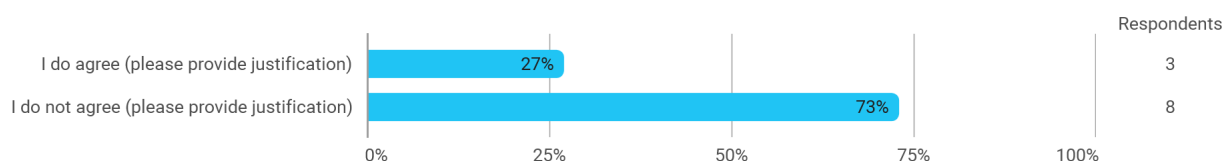
Interested to be part of an OECD Expert Group



The following suggestions have been collected:

1. It might be considered to make the requirement for analytical check for the behaviour of the chemical compulsory. The comment might be relevant for all tests using artificial soil.

Rational: It is mandatory for aquatic organisms (even for acute tests), including sediment dwellers. Indeed, the dissipation of the chemical largely influences the effect, therefore the endpoint get from the test (which is the main aim). The dissipation rate can considerable different in the test media than in the field. Therefore, the use of the endpoints to compare the toxicity of the different chemicals might be a questionable practice. What the endpoint actually express is a combination of the toxicity and the dissipation rate of the chemical in the test media.



Agreement justification:

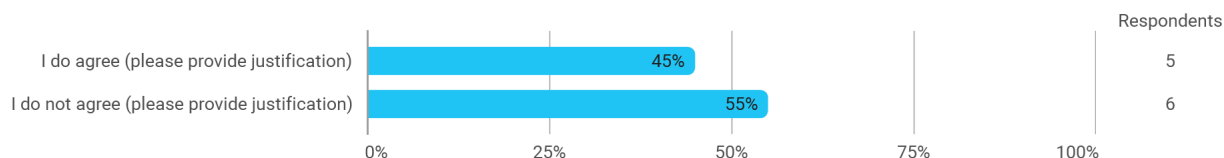
- to confirm the test concentrations

Disagreement justification:

- Behaviour of the test item can be checked with TG 307
- It needs to be defined and explained how and why this should be necessary
- The appropriate guideline for understanding the fate and behaviour of the test substance in soil is OECD 307. Additionally, the use of artificial soil in the Collembola reproduction study would mean such information has little relevance to real world scenarios. However, verifying correct dosing by analysing the dosing solutions would add value to the study.
- General topic to be clarified, not for each individual test guideline
- In case studies are performed under GLP conditions and Certificates of Analysis are provided for each test substance, the amounts of substances or products applied do not need to be analytically verified. Moreover, analytical measurements in GLP soil laboratory tests are in many cases not meaningful due to multiple analytical difficulties when extracting from the soil matrix.
- An inclusion of analytics should be justified by reasonable means. Studies were performed without analytics without any significant issues. Analytics will increase costs with limited benefit. There might be specific test thesis where analytic is helpful. Historically the old approach is sufficient and provides reliable information. Before an inclusion of analytics the use of the information should be discussed also in the context of a potential risk assessment. It should be kept in mind that the compulsory inclusion would lead to a huge repetition of studies unless there are considerations that old studies with no analytics are still valid
- Disagreement. We understand this comment as a suggestion to better check the properties “volatile, unstable or readily degrading substances” in order to judge whether analytical monitoring must be performed. This would in general be a good suggestion, however, we suggest performing analytical monitoring for all substances.

2. New: " Determination of test substance concentration at appropriate time intervals (at least beginning and end of test) " should be included.

Rational: Calculated EC-values will be based on actual concentration instead of the nominal concentration. Such an approach is established for aquatic test systems and it is highly recommended to use a similar approach for soil organisms



Agreement justification:

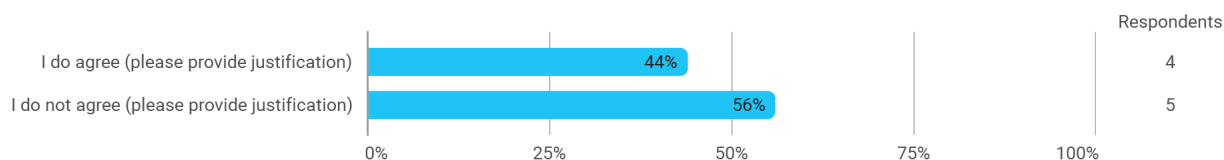
- ▶ Analytical dose verification at the beginning of the test as in many other TG is desirable
- ▶ to confirm the test concentrations

Disagreement justification:

- ▶ Not necessary
 - ▶ The appropriate guideline for understanding the fate and behaviour of the test substance in soil is OECD 307. Additionally, the use of artificial soil in the collembola reproduction study would mean such information has little relevance to real world scenarios. A more appropriate approach is to analytically confirm dosing e.g. by analysis of the dosing solutions. The determination of a mean measured soil concentration or a twa is not appropriate for this study type, because it is not possible to design a study that allows concentrations to be maintained e.g. either a 'flow-through' or a 'concentration renewal' system, to borrow terminology from aquatic studies. Attempting to either 'top-up' soil concentrations or move organisms to a new container has considerable practical problems with a high likelihood of damaging the test organisms and therefore would require ring-testing. The risk assessment is conducted by comparing the peak initial concentration from the exposure models with the effects that are seen at that peak concentration. The current test design reflects the risk assessment requirements.
 - ▶ General topic to be clarified, not for each individual test guideline
 - ▶ In case studies are performed under GLP conditions and Certificates of Analysis are provided for each test substance, the amounts of substances or products applied do not need to be analytically verified. Moreover, analytical measurements in GLP soil laboratory tests are in many cases not meaningful due to multiple analytical difficulties when extracting from the soil matrix.
 - ▶ An analytical verification would significantly increase costs and with very limited additional information. The study is conducted under GLP considerations which provide a sufficient certainty on the correct exposure regime in the soil
3. **New: Guidance for optional experiments designed to calibrate and validate toxicokinetic-toxicodynamic (TKTD) models should be provided. This alternative design should include measurements of organism size, mortality, and reproductive output at intermediate time points (at least 4). Replication at each sampling time is not necessary for parameterisation of TKTD models. Instead replicate test vessels can be sampled at different intermediate time points. For such experiments other procedures could be outlined. Firstly, the experiments begin with 9-12 day old juveniles. Laboratories should be encouraged to gather and/or use relevant data for the growth of younger juveniles (under control conditions) to aid in calibrating the model and more accurately describing the early growth. Further, the variability in ages should be minimised wherever possible. Pulse exposures can be achieved by flooding the test vessels with water, floating the collembola to the surface and allowing transfer to a fresh vessel (Ardestani & van Gestel 2013, Environ. Pollut). The timing of these transfers will have to be managed to minimise potential stress to the organisms. Transfers may also provide opportunities to measure growth, survival and reproduction non-destructively. However, destructive measures are still acceptable.**

References: Ardestani MM & van Gestel CAM. 2013. Using a toxicokinetics approach to explain the effect of soil pH on cadmium bioavailability to *Folsomia candida*. *Environmental pollution*, 180:122-130.

Rational: Toxicokinetic-toxicodynamic modelling, especially when combined with population modelling, can help to quantify and reduce uncertainties in environmental risk assessment by providing improved integration of exposure and effects and by incorporating different pesticide application scenarios and greater ecological realism. TKTD model calibration and validation require variable exposure profiles to demonstrate their ability to extrapolate to untested exposure profiles. Designing experiments with pulsed exposure provides information at the individual level on the timescales of TKTD processes. EFSA Panel on Plant Protection Products their Residues, C. Ockleford, P. Adriaanse, P. Berny, T. Brock, S. Duquesne, S. Grilli, A. F. Hernandez-Jerez, S. H. Bennekou, and M. Klein. 2018. Scientific Opinion on the state of the art of Toxicokinetic/Toxicodynamic (TKTD) effect models for regulatory risk assessment of pesticides for aquatic organisms. 1831-4732.; SCENIHR (Scientific Committee on Emerging and Newly Identified Health Risks), SCHER (Scientific Committee on Health and Environmental Risks), SCCS (Scientific Committee on Consumer Safety). 2012. Addressing the new challenges for risk assessment. Brussels: European Commission.



Agreement justification:

- ▶ Modelling can provide data which can't be generated by experiments in the lab.
- ▶ The suggested updates to these guidelines will enable more realistic and ecologically relevant environmental risk assessments. However, any changes of this type would require ring-testing.
- ▶ However further detailed considerations need to be done to define the needed parameters. In order to not overcomplicate the guideline it might be an option to include the alternative design in an appendix.

Disagreement justification:

- ▶ Not suitable
- ▶ This is a reproduction test
- ▶ I do not agree to include this as a mandatory test design in the main chapters. Modelling is not a useful tool for all substances due to low toxicity of many substances. It must be included as Appendix (optional test design) with the following suggestions: When using non-destructive methods, organism size, mortality and reproductive output can be recorded. Here, camera systems coupled with AI counting systems can be used. These experiments could be conducted on dyed soil (for better contrast) or on dyed plates. While the population densities might be lower on plates, the dynamics are similar. The growth of younger individuals could be assessed on plates in separate experiments. In addition, I agree that the

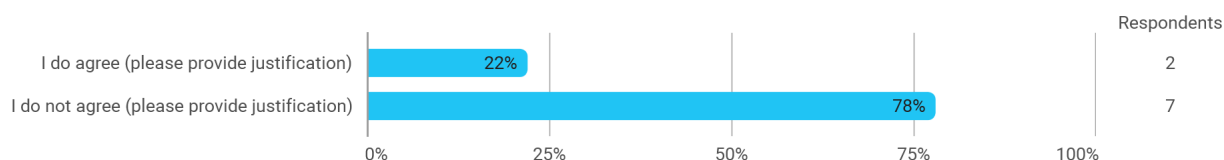
variability in ages should be minimised wherever possible. However, pulse exposure or multi-generation studies should be defined in individual guidelines.

- ▶ Due to high variability and uncertainty associated with the 3-dimensional patterns of toxicant degradation and organism movement in soil (“patchiness”), exposure profiles for soil organisms can be predicted only with high uncertainty. This is already a key issue for Tier 1 experimental testing. Therefore, the use of TKTD models to predict effects on soil organisms in time and space as in the aquatic Tier 2C - where organism exposure in the water column or at the sediment surface is potentially more homogeneous in space- is highly questionable. Developing reliable TKTD models for predicting effects on soil organisms and setting up meaningful scenarios for TKTD modelling in order to identify a realistic worst-case would require huge (even unrealistic) efforts in characterizing the toxicant degradation and the test organism’s behaviour; the species tested should also be suitable to represent of potential toxicologically sensitive and ecologically vulnerable species under realistic field conditions. However, TKTD models could be more successfully applied to retrospectively analyse and understand differences in effects on soil organisms observed across different studies, e. g. across soils having different characteristics. In general, we agree that analytical measurements should be performed in laboratory studies in order to better link exposure to effects. In this regard, residues should be analyzed at different exposure levels in the soil, soil pore water and organisms at the beginning, during and at the end of the study.

4. **para 8 - The following sentence should be deleted: “For volatile, unstable or readily degrading substances (e.g., data generated from a TG 307 study may be considered), or where there is otherwise uncertainty in maintaining the nominal soil concentration, analytical measurements of the exposure concentrations at the beginning, during and at the end of the test should be considered.” Instead, analytical measurements should be performed for all substances, covering the concentration range. Here, the concentration in total soil as well as the concentration in pore water should be determined. Following the EFSA Scientific Opinion on in-soil organisms (2017), “it is recommended that the concentrations of active substances in soils is measured more than twice (see also the EFSA DegT50 EFSA Guidance), depending on the degradation of the active substances and the length of the test.” Suggested new phrasing: “For all substances, analytical measurements of the exposure concentrations at the beginning, during and at the end of the test should be considered. If the concentration of the substance being tested has been satisfactorily maintained within $\pm 20\%$ of the nominal or measured initial concentration throughout the test, analysis of the results can be based on nominal or measured initial values. If the deviation from the nominal or measured initial concentration is not within the range of $\pm 20\%$, analysis of the results should be based on geometric mean concentration during exposure or on models describing the decline of the concentration of the test substance.”**

Rational: According to EFSA Scientific Opinion on in-soil organisms (2017, chapter 7.11.3), analytical measurements for tests on in-soil organisms are requested. In this Scientific Opinion, it is stated that “During laboratory handling procedures of the spiked soils, however, possible losses of the pesticides due to volatilization, degradation, and sorption to, e.g. the glass matrix of vessels used, may occur.” Therefore, the actual concentration of the tested substance needs to be determined in order to correctly link exposure and effects. Moreover, analytical measurements in test systems are included in all aquatic studies. Terrestrial studies should be analyzed

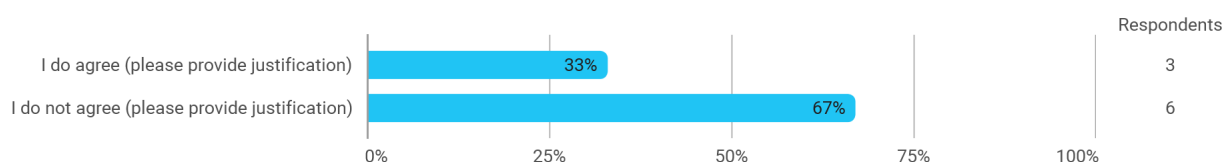
in the same way, otherwise ‘real’ exposure cannot be described and an extrapolation from lab to field is questionable.



Disagreement justification:

- ▶ What is the consequence for the validity of the study? How to handle the control if geometric mean is used for data evaluation.
- ▶ It needs to be defined and explained how and why this should be necessary
- ▶ See points 1 & 2 above.
- ▶ General topic to be clarified, not for each individual test guideline
- ▶ In case studies are performed under GLP conditions and Certificates of Analysis are provided for each test substance, the amounts of substances or products applied do not need to be analytically verified. Moreover, analytical measurements in GLP soil laboratory tests are in many cases not meaningful due to multiple analytical difficulties when extracting from the soil matrix.
- ▶ An inclusion of analytics should be justified by reasonable means. Studies were performed without analytics without any significant issues. Analytics would again increase the costs and complexity and would only provide limited information. There might be specific questions where analytics is justified, however it should remain as an optional point. It should be kept in mind that detection of residues in soil has different complexity than for example in the aquatics where the 80% criteria is used. In general analytics in soil lab studies will lead to a lot of challenges regarding the interpretation.

5. **para 8 - This is a more general question for soil ecotoxicology studies: the data requirements for active substances section 4 (analytical methods) states that: "Methods shall be submitted, with a full description, for the determination of non-isotope-labelled residues in all areas of the dossier, as set out in detail in the following points: in soil, water, sediment, feed and any additional matrices used in support of ecotoxicology studies;" However in practice, analytical methods for soil ecotoxicology studies are almost never provided, since it is not a direct requirement in the guideline. it is still defined as "optional": For volatile, unstable or readily degrading substances (e.g., data generated from a TG 307 study may be considered), or where there is otherwise uncertainty in maintaining the nominal soil concentration, analytical measurements of the exposure concentrations at the beginning, during and at the end of the test should be considered. Would it be possible to clarify that analytical verification of test concentrations in soil is required in the guideline?**



Agreement justification:

- ▶ General topic to be clarified, not for each individual test guideline
- ▶ Total agreement. We made a similar suggestion with point 4.

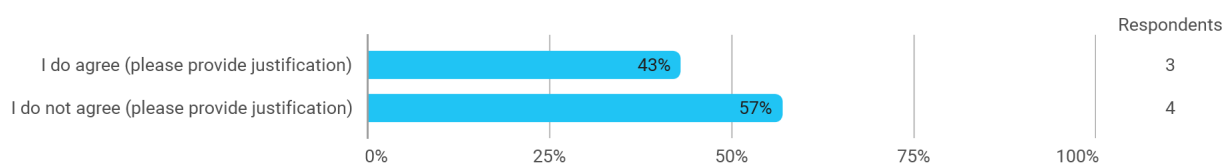
Disagreement justification:

- ▶ Please clarify the analytical requirements first (analytical dose verification, soil extraction, biodegradation...)
- ▶ It needs to be defined and explained how and why this should be necessary
- ▶ See points 1 & 2 above.
- ▶ In case studies are performed under GLP conditions and Certificates of Analysis are provided for each test substance, the amounts of substances or products applied do not need to be analytically verified. Moreover, analytical measurements in GLP soil laboratory tests are in many cases not meaningful due to multiple analytical difficulties when extracting from the soil matrix.

- ▶ see above

6. **para 8 - Further guidance on follow up of the last sentence of this paragraph ('For volatile (....) considered.') is considered necessary. I.e.: criteria to be used for deciding when a substance is considered 'unstable' (e.g. a DT50-value) when measurements are performed, criteria for when to use nominal or measured concentrations (probably the 80% criterium) how to express measured concentrations (i.e. geomean, twa) when no measurements are performed while considered necessary based on the above, what are the consequences for the validity of the study.**

Rational: The 'strictness' of the last sentence of this paragraph ('For volatile (....) considered.'), is not very clear and it leaves open several questions.



Agreement justification:

- ▶ Further guidance for the handling with volatile substances is needed

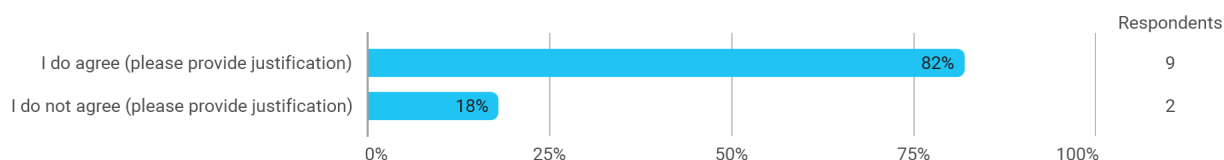
Disagreement justification:

- ▶ See points 1 & 2 above.
- ▶ In case studies are performed under GLP conditions and Certificates of Analysis are provided for each test substance, the amounts of substances or products applied do not need to be analytically verified. Moreover, analytical measurements in GLP soil laboratory tests are in many cases not meaningful due to multiple analytical difficulties when extracting from the soil matrix.
- ▶ Partly agree. However, we suggest performing analytical monitoring for all substances.

7. para 13 - Alternatives to sphagnum peat as a component of OECD artificial soil should be investigated.

Rational: Renewable alternatives to sphagnum peat as a component of OECD artificial soil should be investigated given the environmental concerns associated with peat harvesting. De Silva & Van Gestel (2009) suggest composted coco peat as a suitable replacement. In horticulture, sphagnum peat in potting compost is increasingly being replaced with alternatives such as coir (i.e. coco peat) - how much longer will sphagnum peat be readily available for?

De Silva, P.M.C.S., Van Gestel, C.A.M., 2009. Development of an alternative artificial soil for earthworm toxicity testing in tropical countries. *Appl. Soil Ecol.* 43, 170–174.



Agreement justification:

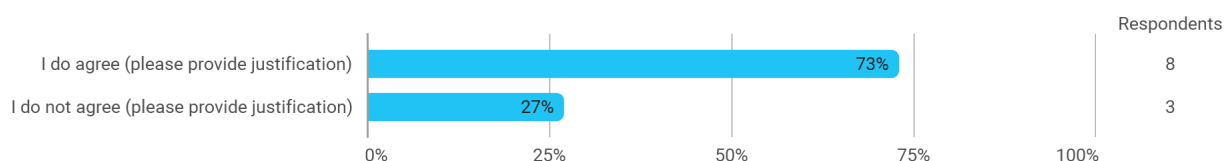
- ▶ Peat moss is not sustainable
- ▶ limited availability of peat
- ▶ In the interests of sustainability alternatives to peat should be actively investigated.
- ▶ sphagnum peat highly variable
- ▶ The extraction of peat contributes directly to climate change. In contrast, using alternatives would contribute to species conservation and climate protection. In the future, the extraction of peat could also be prohibited by law for climate protection reasons. Therefore, it is to be supported to look for alternatives soon.

Disagreement justification:

- ▶ Since the artificial OECD soil provided reliable and reproducible results, I do not see the big need for alternatives. Regardless a search for new alternative might be appropriate but would also need for additional feasibility/ring testing.

8. para 13 - An alternative to sphagnum peat as source of organic matter should be provided.

Rational: The OECD artificial soil requires 5 % organic matter. It is recommended to use sphagnum peat. However, sphagnum peat is not available worldwide and its use is also questionable for ecological reasons.



Agreement justification:

- ▶ Peat moss is not sustainable
- ▶ limited availability of peat

- In the interests of sustainability alternatives to peat should be actively investigated.
- sphagnum peat highly variable
- The extraction of peat contributes directly to climate change. In contrast, using alternatives would contribute to species conservation and climate protection. In the future, the extraction of peat could also be prohibited by law for climate protection reasons. Therefore, it is to be supported to look for alternatives soon.

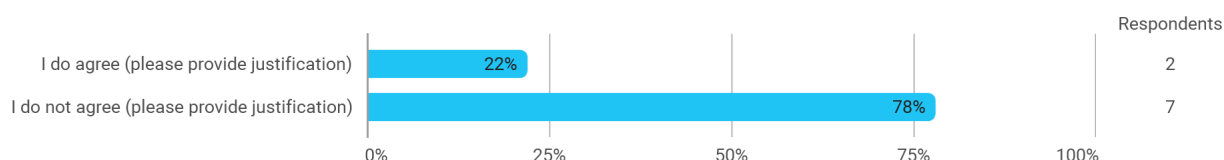
Disagreement justification:

- An alternative to sphagnum peat as source of organic matter can only be provided after a feasible validation and ring-test process.
- Since the artificial OECD soil provided reliable and reproducible results, I do not see the big need for alternatives. Regardless a search for new alternative might be appropriate but would also need for additional feasibility/ring testing.

9. para 22 - The sentence “Food is added after spraying.” Should be deleted. Instead, the food added to each container should be treated with the tested substance at the corresponding concentrations.

Rational: The suggestion follows the “Scientific Opinion addressing the state of the science on risk assessment of plant protection products for in-soil organisms”, EFSA (2017), according to which “exposure via the litter layer is [...] a relevant route of exposure for some soil invertebrates, like macroarthropods, slugs and snails. [...] This may lead to underestimation of internal exposure due to dilution.” An underestimation of effects has been shown by Natal-da-Luz et al. (2019)¹ in a comparison of sensitivity of *Hypoaspis aculeifer* when being fed with contaminated or non-contaminated food in addition to exposure via soil. *H. aculeifer* showed higher sensitivity to the tested substance when fed with contaminated food (mites). These results suggest that oral exposure should be taken into account and could be transferred to the collembolan test system.

1 Natal-da-Luz, T., Gevaert, T., Pereira, C., Alves, D., Arena, M., & Sousa, J. P. (2019). Should oral exposure in *Hypoaspis aculeifer* tests be considered in order to keep them in Tier I test battery for ecological risk assessment of PPPs?. Environmental pollution, 244, 871-876.



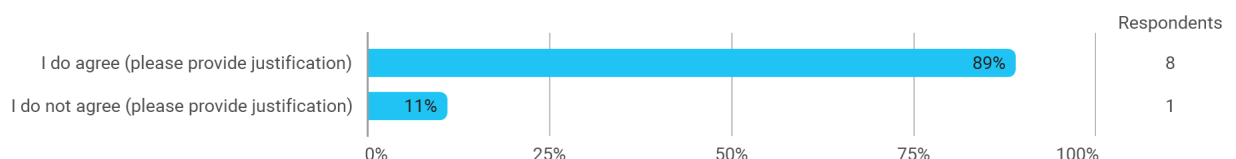
Disagreement justification:

- Under field conditions, the animals would have uncontaminated food available as well.
- Definitely not!
- Exposure via contact and exposure via ingestion are two completely different pathways. Significant research needs to be undertaken prior to introducing a change of this sort. Firstly, to understand the major exposure pathways for collembola and secondly to understand the impact of this change on the test output and subsequently how those data would be

contextualised in a risk assessment. Finally, a change of this kind would require ring testing. It would be more appropriate to consider exposure during feeding in a separate study, so that any toxicity can be attributed to the appropriate exposure pathway and therefore contextualised within the risk assessment. Additionally, dosing the food may be impossible practically for Collembola. They are fed with 2-10 mg of granulated dried baker's yeast at test start and after c. 14 days. Weighing out an appropriate amount of test substance to dose these tiny amounts will be challenging, and furthermore wetting the yeast in order to mix in the test item will activate it, impacting its growth and the frequency with which it should be added to test containers. Lastly, the test chemical itself may directly impact the yeast (fungicidal effect) but not be toxic to Collembola themselves (insecticidal effects) and therefore could confound the purpose of the test.

- ▶ No feeding study
- ▶ In case provided food (yeast) will be contaminated, the food could change in quality (less nutritious due to e.g. Fungicidal or antimicrobial activity) and might not be eaten by the test organism. In this case you would measure an indirect effect (starvation of test organism) and not the direct substance (toxicity) effect.
- ▶ The test system was validated as is. If different exposure scenarios should be included the technical feasibility and reproducibility needs to be checked again. Furthermore accurate application of the food item will be challenging and need to be evaluated in order to gain an understanding on the accuracy/variability and therefore reproducibility

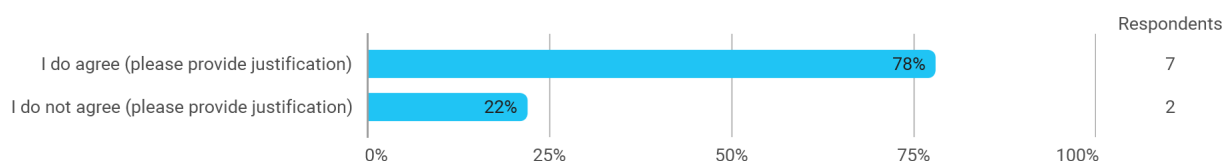
10. para 31 - It is stated that “The validity includes extraction efficiency of juveniles greater than 95%, e.g. by adding a known number to soil.” However, it should be specified how often the extraction efficiency is to be evaluated, e.g. 1-2 times per year. Rational: The guideline is not specific enough on this point.



Agreement justification:

- ▶ Extraction efficacy should be evaluated regularly
- ▶ Once a year
- ▶ Clarification on the expected frequency of extraction efficiency would be beneficial.
- ▶ I do agree. An interval of once per year is sufficient.
- ▶ once a year should be sufficient

11. Annex 3 par. 2: change the temperature range from 20+-1°C to 20+-2°C, change the dark-light cycle from 12-12 hours to 16 hours light and 8 hours dark. Rational: to harmonize the climatic conditions of the culture with the climatic condition of the test



Agreement justification:

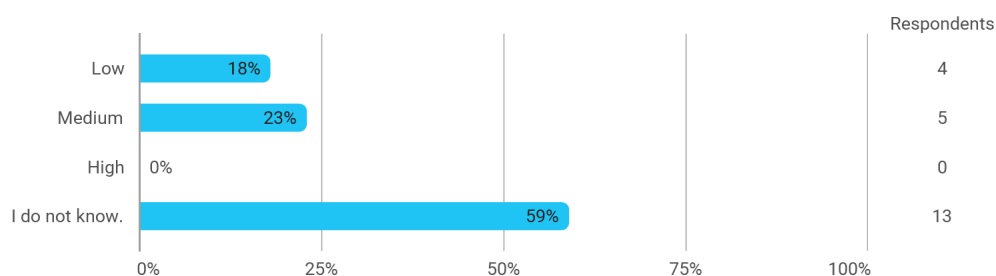
- TG should be harmonised. Furthermore, not only the breeding and synchronisation temperature, but also the test conditions should be adjusted to 20 +- 2°C as in the hypoaspis and earthworm TG.
- more practicable
- however, for the synchronization in general a dark light regime is not necessary for the culture
- Agreement, as conditions during the test and during rearing should be the same.

Disagreement justification:

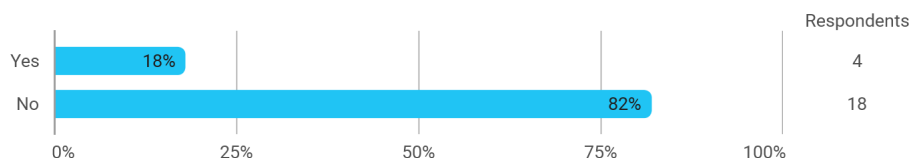
- Changing the light/dark cycle could influence egg laying. A change of this type would require appropriate research and ring testing to understand the impact to the study outcomes.

B.19 Test No. 228: Determination of Developmental Toxicity to Dipteran Dung Flies(*Scathophaga stercoraria* L. (Scathophagidae), *Musca autumnalis* De Geer (Muscidae))

Relevance of revision

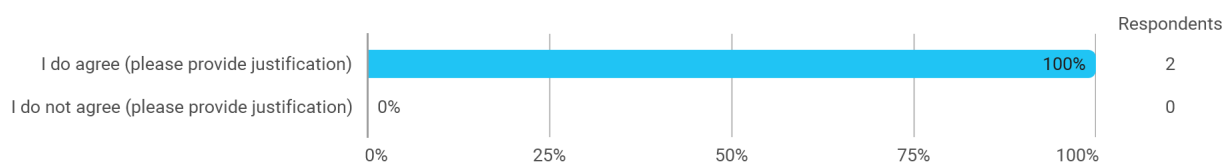


Interested to be part of an OECD Expert Group

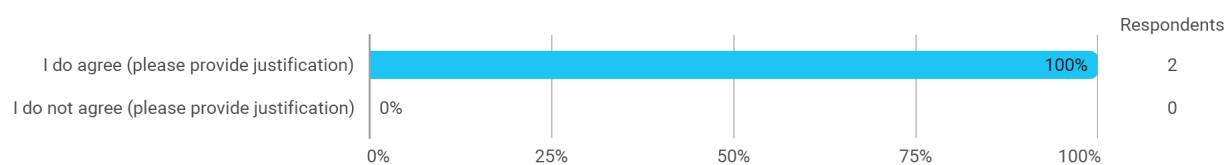


The following suggestions have been collected:

1. para 9 - Question: Why is the timing of emergence (18 ±2 days for *Scathophaga* and 13±2 days for *Musca*) a validity criterion in this test system?



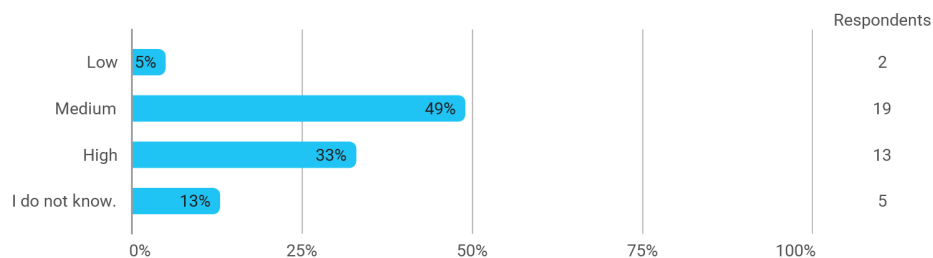
2. para 16 - Change moisture content from "80 % fw" to "80 - 90 % fw".



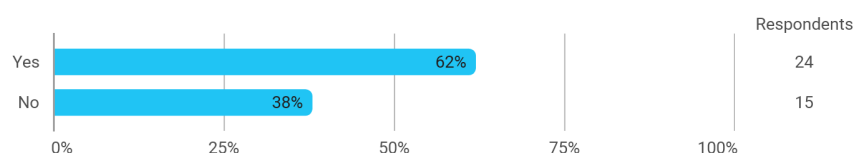
C Assessment feedback Block C

C.1 Test No. 203: Fish, Acute Toxicity Test

Relevance of revision



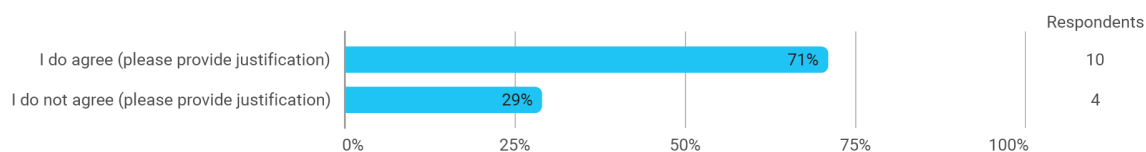
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

- General: Clarification is needed on what type of analytical measurements are compulsory and whether the test is valid even if the analytical measurements indicated very fast dissipation.**

Rational: In cases when the analytical measurements indicated very fast dissipation/complete elimination within the test duration (or no sufficient analytical determinations are available, e.g. no measurement at the end of the test), the validity of the endpoint become questionable; in fact the test is not necessarily measure the toxicity, but the reflects the behaviour of the chemical which is not the aim of the test.



Agreement justifications:

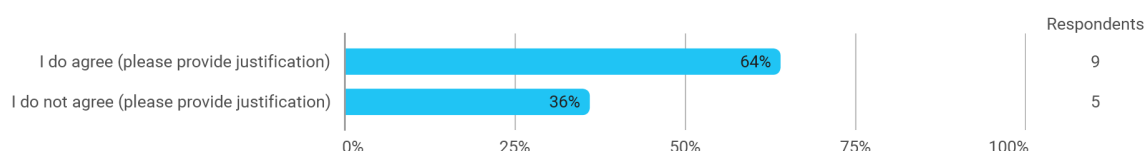
- ▶ some instable substances dissipate very fast and this can cause problems regarding the validity of a test
- ▶ See OECD GD 23 (2019)
- ▶ Not clear how fast dissipation should be considered geomean?
- ▶ Regulators require more analytical criteria than describing in the TG
- ▶ I think it is needed
- ▶ Some additional clarification on compulsory endpoints would be useful, but due to the diverse applications of this assay, it may not be possible to be exhaustive

- ▶ clear direction on when and where analytical measurements are needed is advisable and if not stable then direction on other test methods (flow-through) should be provided
- ▶ The compulsory should be the bare minimum. That is verification of test concentrations. If the analytical chemistry demonstrates rapid dissipation, then further considerations are needed; analytical measurements frequent enough to determine a rate? if so, then could calculate geometric mean concentrations. if not, then would need to terminate and conduct under conditions where stability can be reasonably (app. 60%) maintained or discuss why this is not possible.

Disagreement justifications:

- ▶ There are different reasons for the performance of acute fish studies existing. For hazard characterization stability over test duration is needed. For risk characterization that is not the case. As the acute fish study is a vertebrate study invalidation should be handled with care. The available information resulting from a study should be checked, as well using WoE approaches, to decide whether a study still can be accepted despite problematic stability of the test item. WOE as vertebrate study. If a compound is extremely unstable that would limit its bioavailability under environmental conditions as well. This should be taken into consideration.
- ▶ This can cause repetition of multiple historical studies causing massive increase of animal use.
- ▶ General considerations handled elsewhere in GDs
- ▶ General considerations handled elsewhere in GDs

2. General: Basically, a good harmonization would be needed.



Agreement justifications:

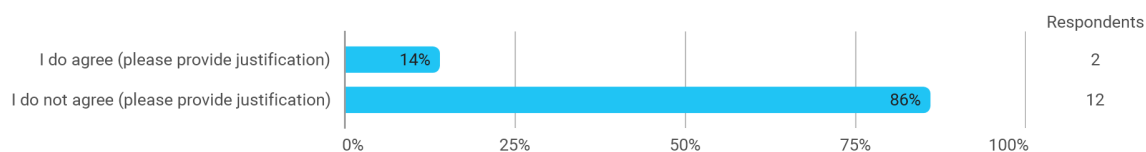
- ▶ harmonization between related test systems would be helpful
- ▶ however the point is too vague.
- ▶ Statement, better harmonization required between TGs
- ▶ harmonization is recommended
- ▶ I think it is needed
- ▶ Agreed, if the discussion is harmonization with the USEPA TGs.

Disagreement justifications:

- ▶ It is not clear what this comment is referring to.
- ▶ Unclear what should be harmonized with what
- ▶ Unclear what "harmonization" is being proposed

- 3. General: The standard mainly deals with use of Zebra fish, but also notes that other fish species such as Guppies can be used without modification of the standard. However, no/limited specific technical data for these other species are available in the current standard.**

Rational: Limited use of Zebra fish in South African laboratories. Guppies is the preferred test organism

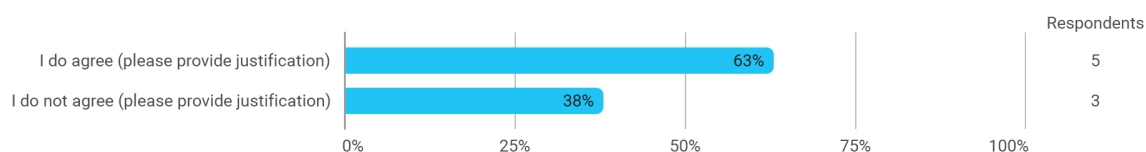


Agreement justifications:

- ▶ More information on test organisms which are not used as standard model for the same test system could be helpful.

Disagreement justifications:

- ▶ The guideline is not specific to zebrafish but covers many species as indicated in the annex.
 - ▶ Guppy is included as a test species in TG 203, Annex 2. Further fish species should not be included without a regulatory need.
 - ▶ Not focus on ZF, seem to be lots of detail for other species in Annex 2.
 - ▶ 11 species of fish along with specific conditions of the test and recommended length range are mentioned in Annex 2. Guppies are specifically mentioned in this Annex. Therefore, the comment is not relevant. In addition, it is not the case that the zebrafish is the main species. In case of pesticides, it is clearly the rainbow trout. And worldwide several species are used for the risk assessment. All needed information is given in Annex 2 of the existing guideline.
 - ▶ Table 1 includes recommended species, it is not specified what details are missing/required
 - ▶ Disagree. This standard does not mainly deal with Zf. Many species are described, with appropriate modifications. However, rationale describing why the species are equivalent would be beneficial
 - ▶ has it been validated on guppies?
 - ▶ Possibly I do not understand the suggestion, but Annex 2 provides species specific information for testing
 - ▶ The species covered are clearly defined - and zebrafish is not the primary species
 - ▶ The species covered are clearly defined - and zebrafish is not the primary species
 - ▶ With all due respect, this TG does not focus primarily on zebrafish. In fact, zebrafish are used on as an example or mentioned 4 times in the TG. The term minnow 5 times. Annex 2 provides enough information on other fish species to conduct an acceptable acute test.
- 4. para 1 - Incorporate an option for testing cationics with and without presence of naturally-occurring dissolved organic substances (i.e. humic acid mitigation) – similar to the US EPA Fish Acute Toxicity Mitigated by Humic Acid draft method.**



Agreement justifications:

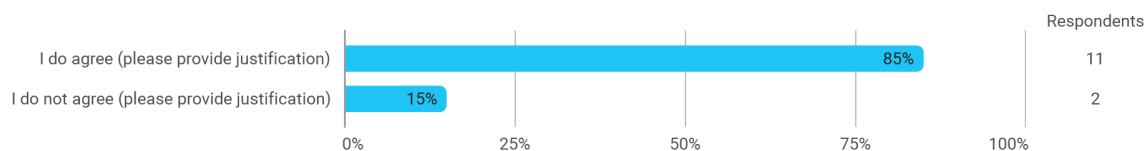
- ▶ however, not sure whether it should be incorporated in paragraph 1
- ▶ I agree, however, this may be more relevant for OECD TG 23 guidance
- ▶ but limit domain

Disagreement justifications:

- ▶ See OECD GD 23 (2019)
- ▶ Before anything can be added to an existing guideline it must be assured that the methodology is valid and could be standardized. If the information on cationics is needed perhaps it could be handled in a separate guideline as handled by US EPA.
- ▶ Unsure if needed, as don't see much ask for this test in pesticide testing

5. para 1 - Add reference to the new OECD TG 249 on the RTgill-W1 cytotoxicity assay: "fish embryos (OECD 2013, fish cell lines (OECD 2021) and others."

Rational: TG 249 is a relevant alternative method for acute fish toxicity. For clarity, TG 249 should be referenced.



Agreement justifications:

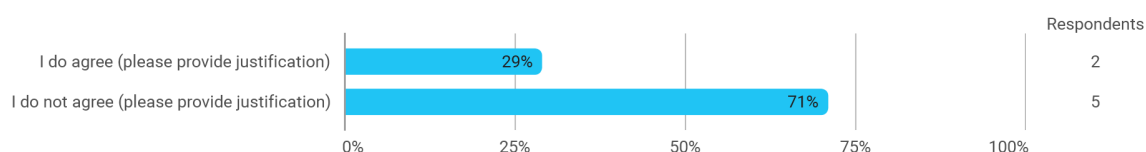
- ▶ animal welfare
- ▶ Should fish embryos (OECD 2013, fish cell lines (OECD 2021) be alternatives?
- ▶ Reference to TG 236 is already there and reference to TG 249 needs to be made. So far, there is only reference to "fish cell lines".
- ▶ update needed
- ▶ Agreed. The RTgill now has an OECD TG and should be mentioned
- ▶ update needed
- ▶ yes, reference could be updated
- ▶ Encouraging alternatives to be considered is good practise
- ▶ These are now accepted alternative methods.

Disagreement justifications:

- The reference was not given as TG 201 from 2019 was older than the new OECD TG 249 RTgill-W1 Guideline. But reference in general to fish cell lines is given in the introduction. In a new version of this guideline the reference should be given. This is not an urgent reason for re-opening the current OECD TG 203 version now.

- para 4 - Please see recent Ctgb comment on CIRCABC by UBA proposed alternative scheme for formulation testing (when waiving the fish oral formulation study is possible):** <https://circabc.europa.eu/ui/group/0b40948d-7247-4819-bbf9-ecca3250d893/forum/topic/c04ab87c-d42e-4832-b015-64441aeffeca>

Rational: Animal welfare reasons/ to follow the principles of RRR where possible



Agreement justifications:

- Agree nor disagree because the page is not available at this time.

Disagreement justifications:

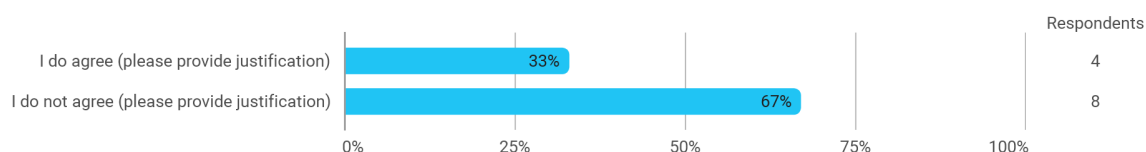
- This comment is about data requirements for a specific type of test item in a specific regulatory framework. An OECD guideline is not the right place for that
- Link is not working
- Detail not relevant
- I can't see the link outcome - not sure I understand what a fish oral test is - more context needed to comment

- para 11 - Add: "A retrospective evaluation of acute toxicity of pesticides showed that the testing of cold water species can be sufficient because they are more sensitive than warmwater or saltwater species for most substances evaluated."**

Rational: The testing of multiple fish species is mostly not required to evaluate the potential acute fish to fish from pesticides. For reasons of animal welfare and efficiency, the number of fish species used for testing should be minimised. In addition to the poster reference below, these analyses will be submitted to a peer-reviewed scientific journal for publication.

Ceger P, Allen DG, Choksi N, Daniel A, Eckel W., Hamm J., Johnson T, Kleinststeuer N, Lowit M, Sprankle C, Truax J, Odenkirchen E (2021) Retrospective Evaluation of the Acute Fish Toxicity Test for Pesticide Registration. Society of Toxicology Annual Meeting, poster, abstract 2324.

<https://ntp.niehs.nih.gov/iccvam/meetings/sot21/ceger-poster-sot2021.pdf>



Agreement justifications:

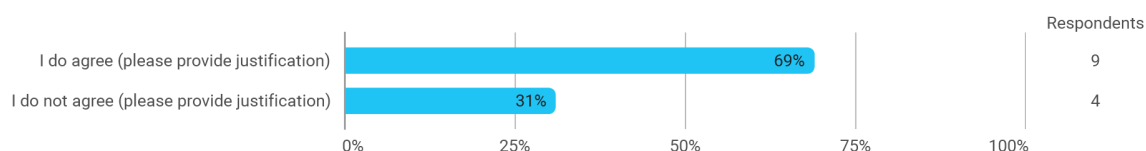
- ▶ animal welfare
- ▶ The proper reference to the analysis of the US EPA OPP presented at the World Congress 11 needs to be made. See the poster of Ceger et al. 2021: Retrospective Evaluation of the Acute Fish Toxicity Test for Pesticide Registration.
- ▶ it would be good to have a positive statement to reduce testing multiple species.

Disagreement justifications:

- ▶ As long as no reference to the cited work is provided that allows to evaluate the relevance of the work this cannot just be taken up. In addition, this comment is again about data requirements and an OECD guideline is not the right place to discuss this topic.
- ▶ this can be added in EFSA document, too specific.
- ▶ Detail not relevant.
- ▶ Can be sufficient for what/whom? I think the information can be given (i.e. coldwater species seems to be more sensitive) and would be relevant/informative, but then what is sufficient should be decided by the regulatory authorities in the geographical area concerned.
- ▶ do not suggest including this at this time as it seems premature at this time due to how multiple regulatory authorities would need to consider it in terms of their testing needs.
- ▶ This is a data/information requirement point out of scope for a TG
- ▶ I doubt regulatory agencies will agree with this unless you can show that this is the case 100% of the time with all chemicals to all fish tested.

8. **para 24 - It would be useful to clarify how to consider the information from the analytical measurements for the interpretation of the test. Note that this comment is relevant for all aquatic tests. However, the different TGs give different level of details on that issue; basically a good harmonization would be needed. Most of my comments for the terrestrial are also general.**

Rational: This paragraph gives some hits how to perform the analytical measurements, but it does not give guidance if and when the endpoint could be expressed as nominal concentrations or when it is recommended to consider the measured values and how.



Agreement justifications:

- ▶ See OECD GD 23 (2019)
- ▶ Same as suggestion 1
- ▶ Also to include clarification on how to report the endpoints if multiple Ais are measured.

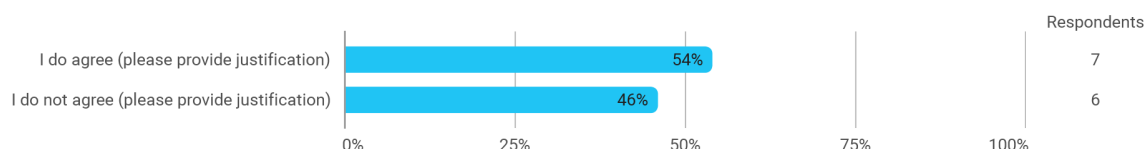
- ▶ I think it is needed
- ▶ Harmonization would be helpful across the aquatic TGs

Disagreement justifications:

- ▶ The regulatory requirements worldwide differ. Therefore, it is difficult to give clear globally accepted guidance in a guideline. Guidelines deliver methods and not the regulatory framework. There are rules from the EU on when to use geometric mean or arithmetic mean or nominal concentrations, but these are not identical with underlying principles e.g. in the US. Therefore, harmonization here would need quite some time. The guideline should focus on how to derive the data. This is the case already. Reference to OECD guideline on difficult substances might help.
- ▶ This is contained within OECD TG 23 guidance, but could be repeated generally here if needed.
- ▶ unsure if needed at this time for the level of effort needed to update, If updated for other scientific reasons then could include.
- ▶ The issue here is not sublethal effects but the determination of morbidity in fish. This should only be considered further if there is a clear way to address this and if this is harmonized globally. For example, morbidity is not currently considered by the USEPA. in their acute fish test guideline. the result of a lack of harmonization would be increased number of acute fish tests and the increased testing of vertebrate species. A global approach is a critical component of this discussion.

9. **Clarity is needed on the consideration of sublethal clinical signs for the endpoint estimation, particularly if a dose-response pattern can be seen. Currently, (some of) these clinical sublethal parameters only need to be reported, but are not taken into consideration any further, even though in certain cases moribundity is evident. Annex 4 of the in 2019 revised OECD 203 guideline provided tools for scientific data collection to generate international consensus on prediction of moribundity via (early) sublethal clinical signs. Is it possible to update the guideline with regard to this topic?**

Rational: Sometimes the reported sublethal clinical signs suggest impending death and must thus be considered an acute effect, but are then not taken into account as there is not clear guidance on how and when to take them into account in the endpoint calculation.



Agreement justifications:

- ▶ moribund fish should not be ignored and be part of the NOEC
- ▶ more clarity would be helpful

- ▶ If evident toxicity is ever to be considered in place of lethality as an endpoint for this test, greater efforts need to be invested in defining, recording and reporting and interpreting the sublethal signs,
- ▶ Should be updated accordingly

Disagreement justifications:

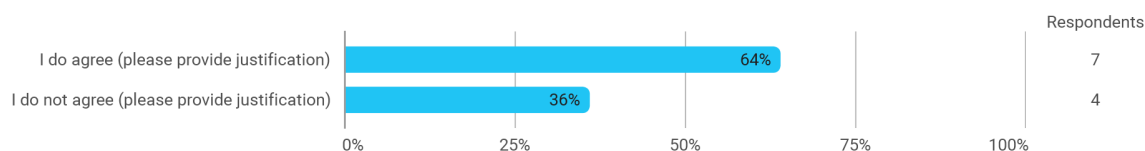
- ▶ was already considered in previous update
- ▶ The inclusion of guidance on the consideration of clinical signs in TG 203 is needed in principle, but the present information on clinical signs has only been included in 2019. Before the TG can be updated again, sufficient data need to be collected to provide evidence for the clinical signs actually leading to mortality.
- ▶ A clear decision on moribund criteria is needed before there is any chance for global acceptability which must be a clear objective of a vertebrate study taking animal welfare into consideration. The scientific community currently is asked to provide the necessary information to allow in the next update of this guideline the setting of defined criteria. As long as this is not the case the undefined moribundity criteria can hardly be used. This will hamper the global acceptance of the studies and might result in an unnecessary increase of vertebrate studies due to non-acceptance of performed studies and finally repetition of studies.
- ▶ Do not suggest going adding additional discussion on morbidity (sublethal endpoints) to predict mortality as there is not clear direction/guidance on this issue yet without further information to support moving away from mortality at this time. Understand efforts may be underway to better understand this and this could be revisited at a later date.
- ▶ This is circular argument whilst the endpoint is defined as mortality

10. para 30 - Include procedures for

- deriving threshold concentration for substances with known instability in water. Should decision be based on nominal or measured based concentrations from alga and daphnia tests?

- UVCBs: same question, and how evaluate multi-component / UVCB mean measured concentrations

Rational: depending on stability mean measured concentration of LC0 may be below EC50 for fish and daphnia. is repetition with another threshold concentration needed (and acceptable with regard to animal welfare? What deviations are accepted. The rate of decomposition of unstable compounds in fish test may be different from daphnia and alga test, i.e. faster and may led to unexpectedly low mean measured concentration being below EC50 values of daphnia and alga.



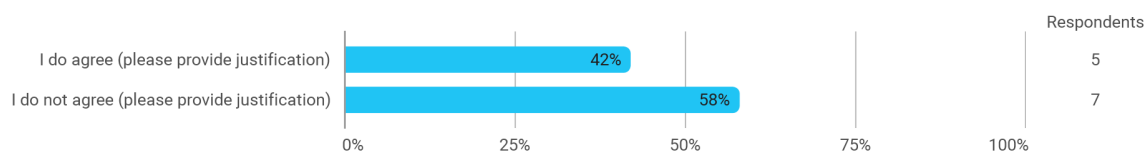
Agreement justifications:

- ▶ more clarity would be helpful
- ▶ but can be done by referencing OECD GD 126 This seems to be too specific to be added to the guideline
- ▶ I think it is needed
- ▶ It would be helpful for the user

Disagreement justifications:

- ▶ See OECD GD 23 (2019). This could be updated if needed.
- ▶ This topic is not necessarily a part of every guideline. Please refer to the OECD Guidance (OECD GD 23) on difficult substances.
- ▶ This is OECD TG 23 guidance
- ▶ I agree with the current guidance for expressing endpoints using nominal or measured test concentration. For unstable compounds, the issue is whether enough sampling was conducted to construct a degradation curve that would allow for the calculation of the test material during the study. Data should not be taken from algal or daphnid studies to apply degradation times for fish. Test conditions (temp) can be very different, as would be adsorption potential (bioavailability) component would be very different. This would be comparing apples to oranges...

11. para 30 - how to deal with missing data for alga and daphnia. Instructions needed here. Are in silico values an acceptable surrogate? Should these tests be carried out first (obligatory?)



Agreement justifications:

- ▶ I believe actual testing of Daphnia and algae would be required to accurately determine the threshold concentration.
- ▶ I think it is needed
- ▶ it would be helpful for the user; however this has to be based on previous work

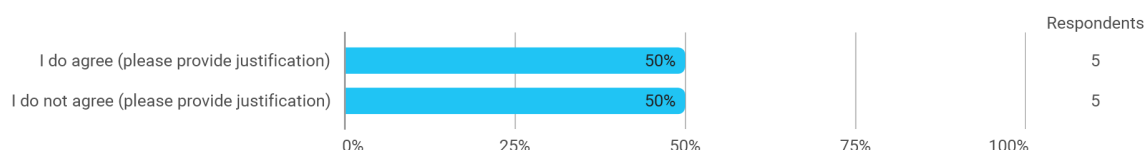
Disagreement justifications:

- ▶ Typically algae and daphnia data are available in most cases.
- ▶ The threshold approach is not obligatory. Information about replacing these tests with in silico methods is specified in the legislation and in guidance documents.
- ▶ Limit Test is well described in chapter 30. The question on which study to be used is not to be clarified in a guideline but in the data requirements. The question for the performance of

an acute fish test should be: Is it really needed for risk assessments, or do we have better possibilities nowadays. That needs to be clarified in general and is not part of this guideline.

- ▶ para 30 does not discuss the threshold approach, the discussion on the acceptance of silico values is too specific.
- ▶ Not relevant for OECD 203 guidance
- ▶ In silico data are not acceptable alternatives at this time. Believe there is sufficient information on how to set limit concentration in this TG and other GD, and if no other data are available then rangefinder may need to be conducted.
- ▶ This is not the place to talk strategically about a product. This is the test guideline to conduct an acute fish test. Algal and daphnid issues are not relevant to this test guideline.

12. The methodology as described in the current guideline is not suitable to test for effects of virtually insoluble particulate materials (e.g. nano-forms of substances, polymers). Adaption of the guidelines methodology with latest scientific findings is indicated so that valid tests with particulate materials can be performed.



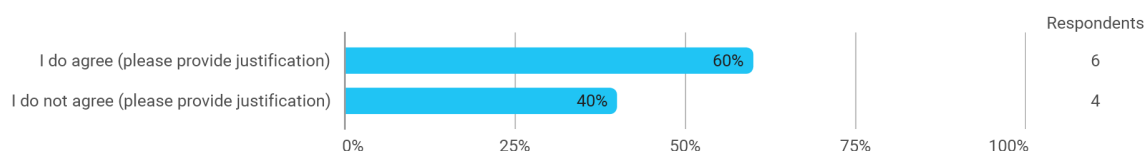
Agreement justifications:

- ▶ More supporting information and suggestions would be helpful in the TG
- ▶ While I agree with the comment, it is ultimately the responsibility of the submitter to report the bioavailable component. The specific may not be applicable due to the number of situations, but handled individually with the product regulatory reviewer.

Disagreement justifications:

- ▶ See OECD GD 23 (2019) and OECD GD for aquatic toxicity testing of manufactured nanomaterials
- ▶ The test is standardized and validated for dissolved test items. For difficult substances please refer to the OECD Guidance on difficult substances (OECD GD 23).
- ▶ This is OECD TG 23 guidance and is outside the scope of current TG 203 guidance
- ▶ Not sure if such an update (which would be needed) could be done so easily, without additional validation work?
- ▶ Believe that sufficient information on this is available in GD23

13. Update the controls text to reflect OECD WNT project 2.55 – Part 2 (Use and analysis of control fish in toxicity studies).



Agreement justifications:

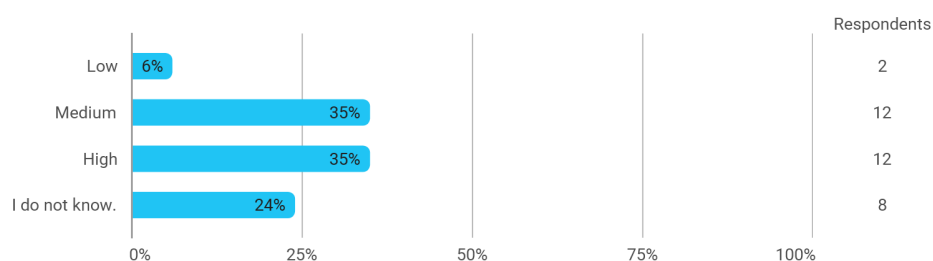
- ▶ state of art
- ▶ 3Rs
- ▶ update needed
- ▶ If guidance is formally accepted, it should be incorporated into an updated 203 TG
- ▶ update needed

Disagreement justifications:

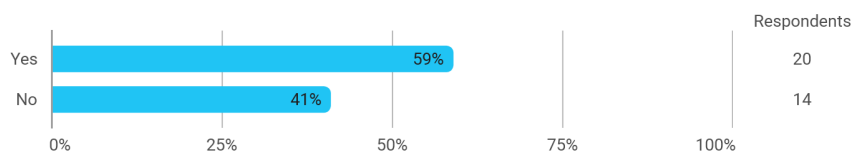
- ▶ Wording on the use of one control is included in paragraph 23 on controls.
- ▶ This para can be updated in the routine refreshing of the guideline and is not urgent. The current guidance already allows the use of a single control.
- ▶ This project seems to be not finished yet.
- ▶ Do not suggest modifying the control language at this time as this is still an topic area for discussion that needs to be considered more fully across members before revisions.

C.2 Test No. 210: Fish, Early-life Stage Toxicity Test

Relevance of revision



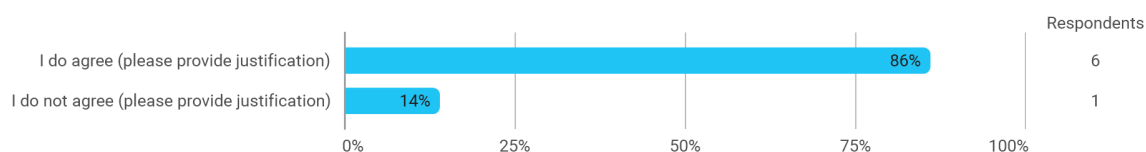
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. Include a section on initial considerations in line with the latest update of OECD TG 203 (18 June 2019)

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

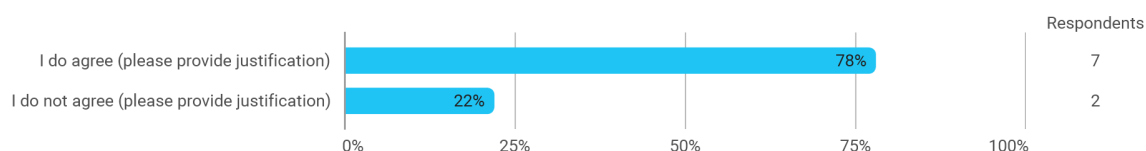
- It is a good idea to have a paragraph on initial considerations in all TGs.
- most up to date information
- All references to TG 203 should reflect any update made to the 203
- As long as the wording is consistent

Disagreement justifications:

- This alone is not a reason for re-opening this guideline. If a new version is coming this should be added, but not by copying from OECD TG 203. Fish Egg Tests and cell line studies do not yet allow the appropriate prediction of chronic effects. Some progress is available on growth related effects. The current introduction of the FELS guideline is therefore still correct. Methods as fish egg tests already can be used as RF testing possibilities. Therefore, an adaptation of the current guideline version is not urgently needed.

2. **para 9 - Add the following text from OECD TG 203: When testing difficult test chemicals, Guidance Document No. 23 (OECD, 2019) should be consulted and the study design modified appropriately. For volatile and other difficult test chemicals, further specific measures should be taken (OECD, 2018). Any silicone materials that were in contact with the test solution(s) should be preferably discarded and not reused for subsequent tests with different test chemicals.**

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.

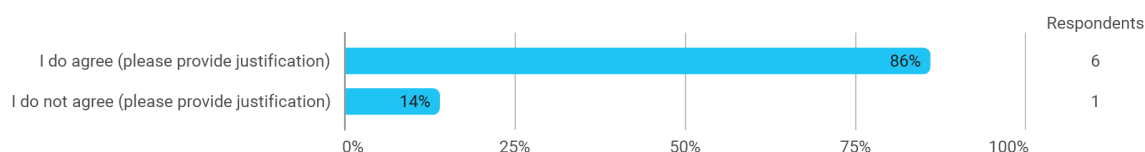


Agreement justifications:

- to clarify difficult chemicals for testing
- Good to reference guidance document 23
- All references to TG 203 should reflect any update made to the 203
- Update needed for harmonization
- As long as the wording is consistent

Disagreement justifications:

- ▶ i think it is not feasible to not reuse all silicone material, since glass tanks are often sealed with silicone. A change of whole tank systems is not possible in some cases. However i agree with not reusing silicone tubes or other single use products.
 - ▶ The reference to OECD TG 23 can be included in the future but is currently no reason for re-opening this guideline.
3. **para 14 - Align the wording with OECD TG 203. OECD TG 211 states that measuring TOC/DOC is only recommended while TG 203 states it should be measured. It is proposed to use the following wording from OECD TG 203: If natural water is used, DOC or TOC and nitrate-content (NO₃) should be measured once prior to the test. Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.**

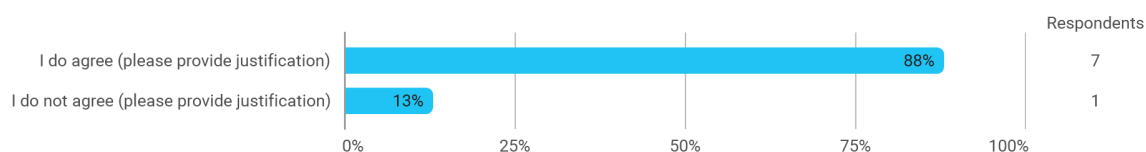


Agreement justifications:

- ▶ align with 203
- ▶ All references to TG 203 should reflect any update made to the 203
- ▶ update needed for harmonization
- ▶ The wording should be consistently. Semantically, they say the same thing.

Disagreement justifications:

- ▶ CROs are doing periodic checks of the DOC/TOC and nitrate but not before each test. Thus, this could be added but is not urgently needed.
4. **para 16 - To be replaced by Para. 17 from OECD TG 203. The following text from Para. 23 of OECD TG 203 should also be included: Low toxicity solvents only (i.e. acetone, ethanol, methanol, tertiary-butyl alcohol, acetonitrile, dimethyl formamide, dimethyl sulfoxide, and triethylene glycol) as recommended in Guidance Document No. 23 (OECD, 2019) should be used whilst solvents of unknown toxicity should not be used. It should be noted that in spite of the low toxicity for fish, dimethyl formamide and dimethyl sulfoxide should be avoided where possible on human health and safety grounds. Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines**



Agreement justifications:

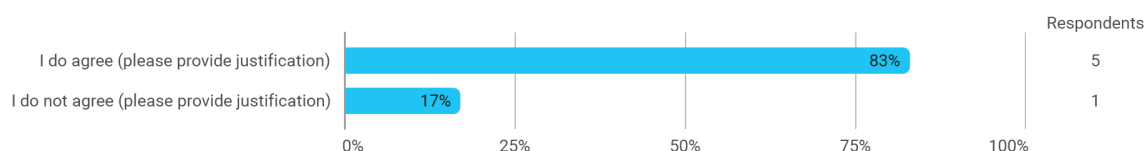
- ▶ Agree. Information about the use of appropriate personal protective equipment could be added.
- ▶ align with 203
- ▶ All references to TG 203 should reflect any update made to the 203
- ▶ update needed for harmonization
- ▶ I am okay with this.

Disagreement justifications:

- ▶ DMF is a routinely used solvent in aquatic testing. It has been shown to be more suitable in most cases as e.g. acetone or ethanol. In general, untested substances, as routinely tested in ecotox labs, have to be handled as if they were CMR compounds. Thus at least the safety measures in the labs should always be appropriate. Based on the fact that DMF was predominantly used most of the historical data being available for possible effects of solvents, relate to DMF. The currently ongoing project investigating whether always a solvent control and a control is needed when a solvent is used in a study based their evaluations as well on a high number of studies which have been performed with DMF. The hope is, that this evaluation will allow to skip the negative control and by that reduce animal numbers in vertebrate tests. If we lose DMF that might have a negative impact on this decision.

5. Add Para. 18 from OECD TG 203.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines



Agreement justifications:

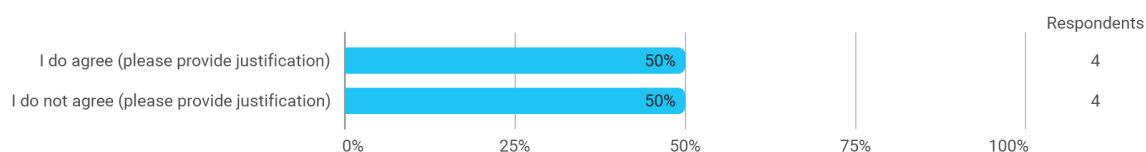
- ▶ align with 203
- ▶ All references to TG 203 should reflect any update made to the 203
- ▶ Reasonable

Disagreement justifications:

- ▶ The TG 203 para 18 calls for performing a test in parallel without pH adjustment if the pH of the stock solution must be adjusted to 6.0-8.5. That is impractical for a chronic fish study. Considering the fact, that unnecessary vertebrate testing needs to be avoided this should not be considered for the FELS test. The pH adjustment in the range of 6 to 8.5 is allowed due to historical findings which demonstrated that this is an acceptable pH range for the used fish species. Thus, in case a pH adjustment is needed there is already sufficient information to consider this and the use of additional fish to prove it again and again is not needed and should clearly be avoided.

6. Frequency of analytical determinations and measurements (Para. 24-25): align with Para 24-25 of OECD TG 203.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines



Agreement justifications:

- ▶ align with 203
- ▶ All references to TG 203 should reflect any update made to the 203
- ▶ update needed

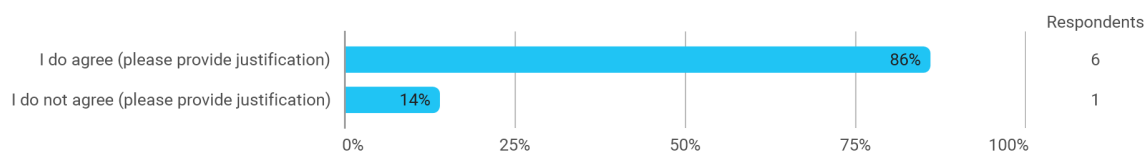
Disagreement justifications:

- ▶ Disagree to align with paragraph 25 about the frequency of measurement of pH, O2, salinity, etc. It is not necessary to measure them daily in all test vessels. Therefore, the para should stay unchanged. With respect to pH, O2, salinity etc. The currently presented information is sufficient in most cases. Therefore, the adaptation of this para is not urgently needed. Agree to align with OECD 203 for para 24 regarding chemical analysis frequency.
- ▶ it is specific enough
- ▶ Measurement of pH oxygen and temperature on a daily base in a flow through system is redundant especially with regards to the duration of the test for above 30 days. measurements should be done once a week, and if necessary due to the substance or other factors increased in a case by case decision
- ▶ TG 210 is very specific when it comes to the frequency of analytical determinations that are not included or could be aligned with the acute fish test (TG 203). I would be open to harmonizing elements from TG203 into TG210.

7. Treatment of results: Add the following text from Para. 31 of OECD TG 203: It is recommended that results should be calculated using the measured concentrations of the test chemical. If the deviation from the nominal concentrations is smaller than

20%, results may also be based on the nominal concentrations. It should be noted that it is often useful to have both measured and nominal effect concentrations quoted, see Guidance Document No. 23 (OECD, 2019).

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines



Agreement justifications:

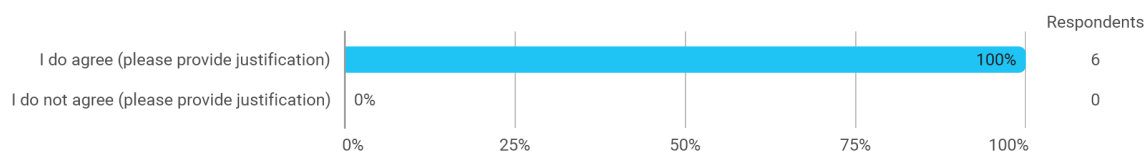
- ▶ align with 203
- ▶ All references to TG 203 should reflect any update made to the 203
- ▶ update needed
- ▶ Whatever wording or approach is taken needs to be consistent with all aquatic and terrestrial tests where analytical measurements are taken.

Disagreement justifications:

- ▶ Disagree. The information given in para 24 is sufficient. Different regulatory frameworks use different approaches for the calculations. Harmonization would be appreciated but might not be the right way to do it via the OECD technical guidelines.

8. Align Annex 4 with the Annex 3 from OECD TG 203.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines



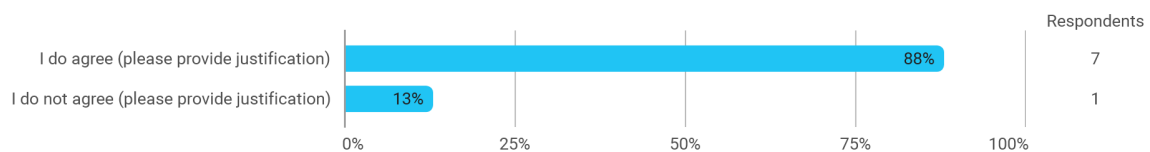
Agreement justifications:

- ▶ align with 203
- ▶ All references to TG 203 should reflect any update made to the 203
- ▶ needed for harmonization
- ▶ Makes sense

9. para 24 - The underlined text needs to be revised: “When the measured concentrations do not remain within 80-120% of the nominal concentration, the effect concentrations should be determined and expressed relative to the arithmetic mean concentration for flow-through tests (see Annex 6 of the Test Guideline 211 for the calculation of the arithmetic mean (8)), and expressed relative to the geometric mean of the measured concentrations for semi-static tests (see Chapter 5 in the OECD Guidance Document No. 23 on aquatic toxicity testing of difficult substances and mixtures (2))”

Rational: The underlined part is not relevant as the calculation method refers to the time-weighted means for a semi-static design. At the same time the OECD GD 54 in Annex 2 provides a method for flow-through exposure that seems more suitable for semi-static exposure.

The OECD GD 23 was recently updated and the order of chapters has been changed, therefore, reference needs to be updated.



Agreement justifications:

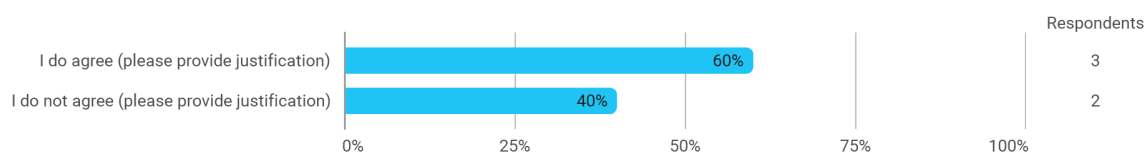
- ▶ to be better used for the regulatory purposes
- ▶ align to other TGs
- ▶ Edits are needed for clarity
- ▶ update needed
- ▶ Makes sense. The wording or approach is taken needs to be consistent with all aquatic and terrestrial tests where analytical measurements are taken.

Disagreement justifications:

- ▶ Different regulatory frameworks use different approaches for the calculations. Harmonization would be appreciated but might not be the right way to do it via the OECD technical guidelines. In general agreement with OECD GD 23.

10. para 27 - “Hatching and survival: observations on hatching and survival should be made at least once daily and numbers recorded”, for practical reasons this could be changed to something like: “Numbers of hatched larvae should be recorded at least daily until completions. Observations on survival should be made at least once daily and numbers recorded. Survival can be assessed by observing mortality during the exposure period. The total survival is assessed at the end of the exposure by counting surviving organisms”.

Rational: Daily counting of surviving larvae is a very laborious and prone to mistake task as the larvae are very small and, towards the end of exposure, very fast and mobile. Taking photos is not an easy option as it is difficult to get a sharp image of the entire volume of exposure vessel. For practical reasons, many labs employ this approach, but it is a deviation to the guideline.



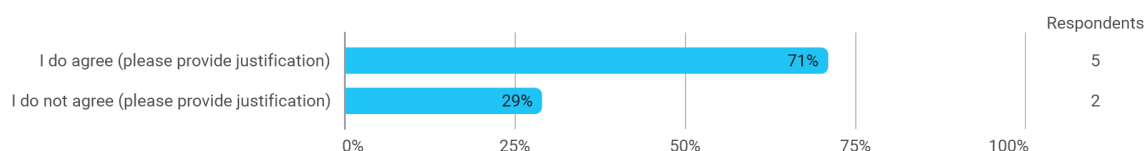
Agreement justifications:

- Edits are needed for clarity

Disagreement justifications:

- Disagree. The current guideline text is more appropriate as the proposal for change.
- The wording is appropriate because daily observations are a must for this test.

11. paras 29, 32, 33 - Clarity is needed on the consideration of sublethal clinical signs for the endpoint estimation, particularly if a dose-response pattern can be seen. Currently, (some of) these clinical sublethal parameters only need to be reported, but it is not clear and when to take them into consideration for endpoint estimation, even though in certain cases moribundity is evident. Rational: Sometimes the reported sublethal clinical signs suggest impending death and must thus be considered an acute effect, but are then not taken into account as there is not clear guidance on how and when to take them into account in the endpoint calculation.



Agreement justifications:

- Relevant clinical signs predicting mortality should be identified also in this long term test, in addition to TG 203.
- align with TG 203
- update needed

Disagreement justifications:

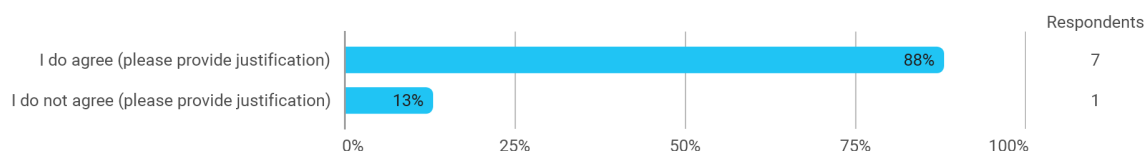
- Disagree: The description of morphological abnormalities and behavioural effects are part of a study report already. No need for a change. Before moribundity criteria can be added to this guideline they need to be defined. That is yet not even the case for the OECD TG 203. The scientific community is asked to deliver information that will allow the definition of moribund criteria. When those will be available is unclear. For the time being the information available is not sufficient to add that to the OECD TG 210 or other fish guidelines. The objective of the OECD TG 210 is to observe non-lethal effects. Therefore, the test concentrations should be chosen to avoid lethal effect as far as possible. The importance of moribund criteria therefore is lower compared to the acute fish tests. RF test with fish embryos might support the correct choosing of test concentrations by delivering information on strong effects on fish embryos. Endpoint derivation on behavioural effects is difficult as

described already in the current guideline version. In cases where a clear dose response is established it might be possible.

- This becomes an issue of global harmonization as much as anything. For example, if morbidity is not harmonized between the EU and the nUSEPAQ, the result will be extra tests and the use of addition, really excessive, numbers of vertebrate species. This is a non-starter if harmonization cannot be attained.

12. Light periods are provided without providing desirable light intensity values

Rational: Such values are provided for all other guidelines.



Agreement justifications:

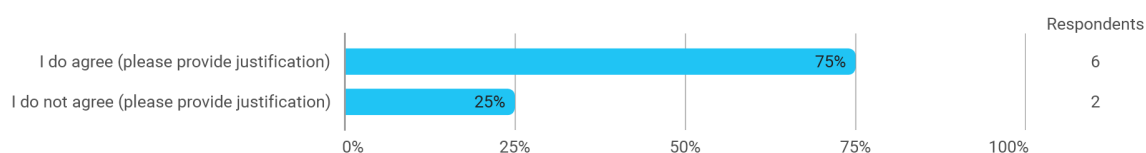
- yes light intensity range would be helpful
- more clarification
- Could be adjusted to 540-1080 lux to be in line with other TG's (OECD 234) for fish testing
- light intensity values could be added
- Units are needed

Disagreement justifications:

- This can be variable and lab dependent.

13. Annex 4 of the in 2019 revised OECD 203 guideline provided tools for scientific data collection to generate international consensus on prediction of moribundity via (early) sublethal clinical signs. Is it possible to update the guidelines 203 and 210 with regard to this topic?

Rational: Sometimes the reported sublethal clinical signs suggest impending death and must thus be considered an acute effect, but are then not taken into account as there is not clear guidance on how and when to take them into account in the endpoint calculation.



Agreement justifications:

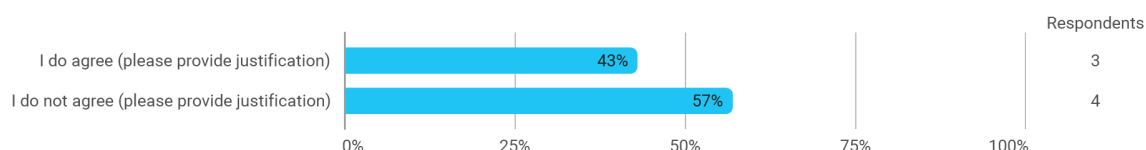
- TG 203 is probably not ready for another update after the last update of 2019, see above. However, in TG 210, relevant clinical signs predicting mortality should be identified also in this long term test, in addition to TG 203.
- align with 203
- clarification on moribundity and behaviours could harmonize the evaluation

- update needed

Disagreement justifications:

- The objective of the OECD TG 210 is to observe non-lethal effects. Therefore, the test concentrations should be chosen to avoid lethal effect as far as possible. The importance of moribund criteria therefore is lower compared to the acute fish tests. No urgency at the moment. Before moribundity criteria can be added to this guideline they need to be defined. That is yet not even the case for the OECD TG 203. The scientific community is asked to deliver information that will allow the definition of moribund criteria. When those will be available is unclear. For the time being the information available is not sufficient to add that to the OECD TG 210 or other fish guidelines.
- Same as Suggestion 11.

14. The methodology as described in the current guideline is not suitable to test for effects of virtually insoluble particulate materials (e.g. nano-forms of substances, polymers). Adaption of the guidelines methodology with latest scientific findings is indicated so that valid tests with particulate materials can be performed.



Agreement justifications:

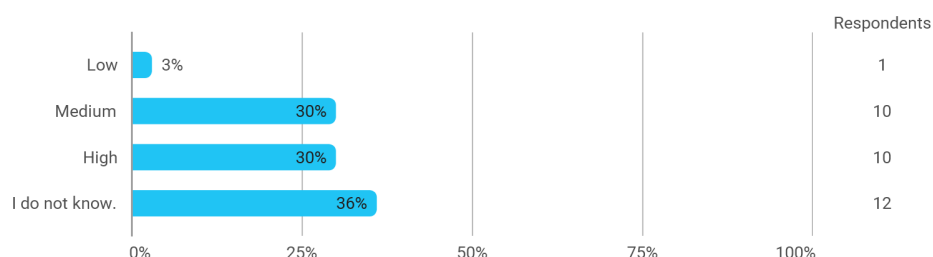
- reference guidance document 23 and nano guidance document
- While I agree with the comment, it is ultimately the responsibility of the submitter to report findings based on the bioavailable fraction. This becomes an issue of what is an acceptable practice in that particular industry and by that particular regulatory agency. i am not sure that the TG is the place to address this, but to instead to acknowledge the point.

Disagreement justifications:

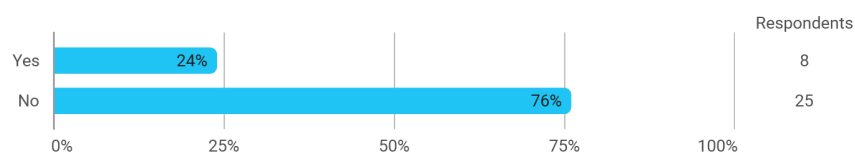
- see OECD GD 23 and OECD GD for aquatic toxicity testing of manufactured nanomaterials
- The test was standardized and validated for dissolved test items. For difficult substances please refer to the OECD Technical Guidance on difficult substances (OECD GD 23).
- This should be a separate guidance
- This guidance belongs in TG23

C.3 Test No. 212: Fish, Short-term Toxicity Test on Embryo and Sac-Fry Stages

Relevance of revision



Interest in being part of an OECD Expert Group



The following suggestions have been collected:

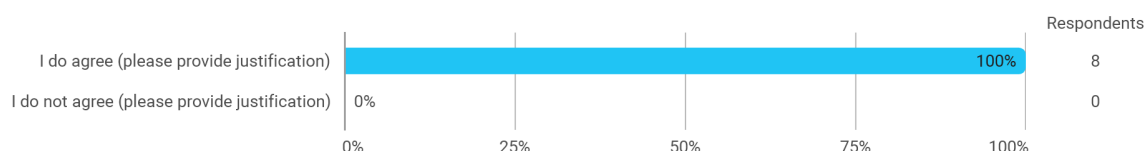
1. Deletion of TG 212 is requested.

Rational: In TG 212, there is a risk that sac-fry stages starve to death, which is a considerable animal-welfare issue. Moreover, TG 212 is less sensitive than TG 210 and restricted to specific log Kow ranges.

OECD Guidance Document 171, Fish Toxicity Testing Framework, paragraph 169

UBA, 2015: REACH Compliance: Data availability of REACH registrations. P. 118f.

https://www.umweltbundesamt.de/sites/default/files/medien/378/publikationen/te_xte_43_2015_reach_compliance_data_availability_of_reach_registrations_0.pdf



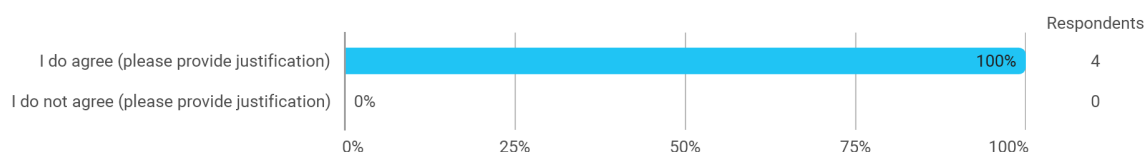
Agreement justifications:

- In TG 212, there is a risk that sac-fry stages may starve to death, which is a considerable animal-welfare issue.
- Not used, seems like a pointless TG
- Agreed. This test is not needed anymore. It is rarely used and delivers by far less information as a FELS study by using similar number of vertebrates. For animal welfare reasons a FELS test is thus the by far better test system.
- Deletion would be justified.
- Need to check if the situation in Japan has changed. From memory this was deleted following the Fish testing framework discussions as there were uses in Japan. On animal welfare grounds it would be good to remove this TG...

2. Table in Annex 3

For Zebrafish, the exposure time should be reduced. Experiments showed that the duration indicated in the OECD guideline is not appropriate. 5 to 6 days corresponds to

normal larvae development (as described in more recent publication on zebrafish). This duration is also compliant with Directive 2010/63/EU (of the European Parliament and of the council of 22 September 2010) on the protection of animals used for scientific purposes.

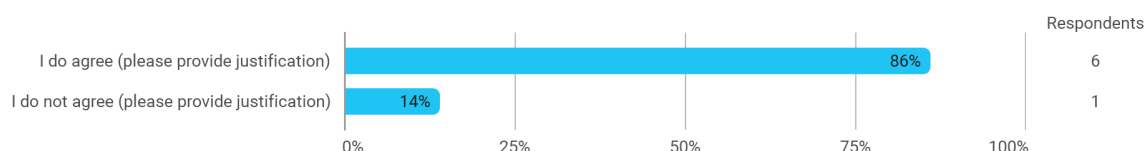


Agreement justifications:

- Annex 1, paragraph 9 of TG 212 says, „At 25 °C the eggs will hatch 3-5 days after fertilization and the yolk sac will be absorbed approx. 13 days after fertilization.“ However, independent feeding and thus protection by Directive 2010/63/EU starts in zebrafish at 120 hpf.

- Technically sound

3. Consider this method as a candidate for 3Rs modification/deletion noting there may be method modifications that could deliver an ethically defensible method which has 3Rs benefits i.e. considered/used as a (sub)chronic screen for in vivo testing.



Agreement justifications:

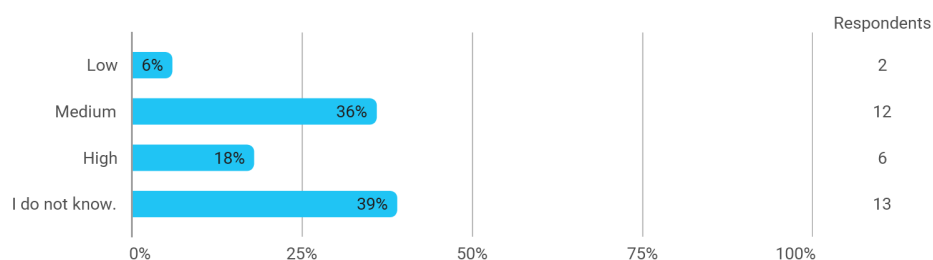
- Other current guidelines are preferred.
- This requires discussion. Would this involve feeding and prolonged exposure?
- see above

Disagreement justifications:

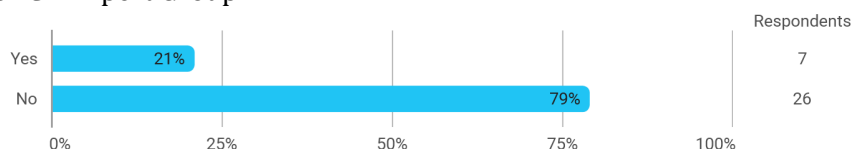
- While I do agree that this method is appropriate for screening purposes, there needs to be a lot more data available to show that this test can replace the in vivo studies. Currently, this is not the case.

C.4 Test No. 215: Fish, Juvenile Growth Test

Relevance of revision



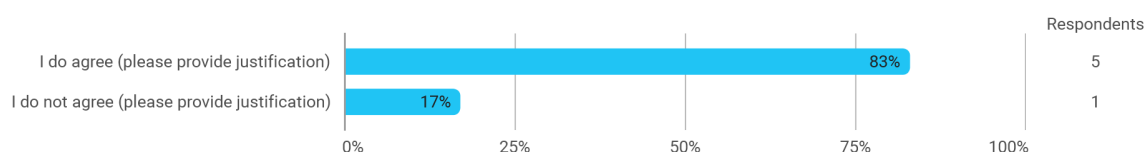
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. **Include a section on initial considerations in line with the latest update of OECD TG 203 (18 June 2019).**

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

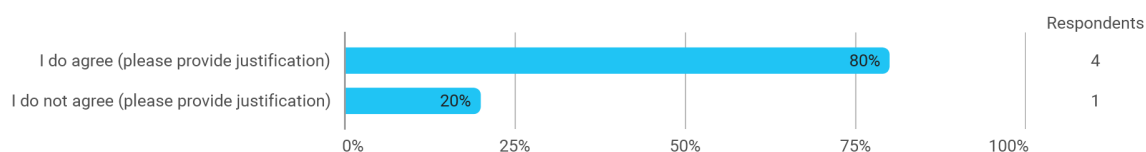
- Agree. It is a good idea to have a paragraph on initial considerations in all TGs.
- Harmonization needed
- Makes sense

Disagreement justifications:

- This alone is not a reason for re-opening this guideline. If a new version is coming this should be added, but not by copying from OECD TG 203. Fish Egg Tests and cell line studies do not yet allow the appropriate prediction of chronic effects. Some progress is available on growth related effects. The current introduction of the FELS guideline is therefore still correct. Methods as fish egg tests already can be used as RF testing possibilities. Therefore, an adaptation of the current guideline version is not urgently needed.

2. **para. 6. Add the following text from OECD TG 203: When testing difficult test chemicals, Guidance Document No. 23 (OECD, 2019) should be consulted and the study design modified appropriately. For volatile and other difficult test chemicals, further specific measures should be taken (OECD, 2018). Any silicone materials that were in contact with the test solution(s) should be preferably discarded and not reused for subsequent tests with different test chemicals.**

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

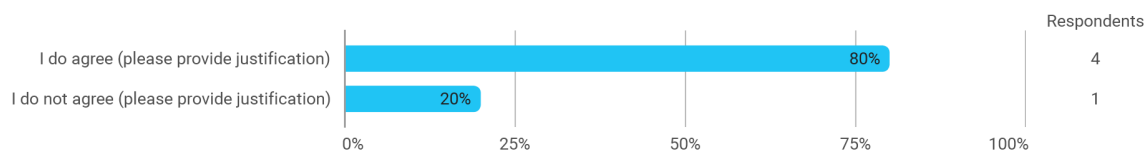
- Harmonization needed
- Makes sense

Disagreement justifications:

- The reference to OECD TG 23 can be included in the future but is currently no reason for re-opening this guideline.

3. **para. 7: align the wording with OECD TG 203 on measuring TOC/DOC. It is proposed to use the following wording from OECD TG 203: If natural water is used, DOC or TOC and nitrate-content (NO₃) should be measured once prior to the test.**

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

- Harmonization needed
- Makes sense

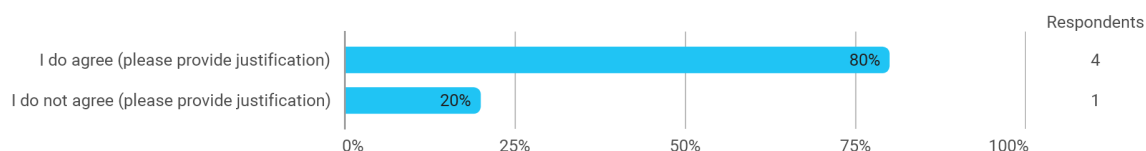
Disagreement justifications:

- CROs are doing periodic checks of the DOC/TOC and nitrate but not before each test. Thus this could be added but is not urgently needed.

4. **para. 8: to be replaced by Para. 17 from OECD TG 203. The following text from Para. 23 of OECD TG 203 should also be included: Low toxicity solvents only (i.e. acetone, ethanol, methanol, tertiary-butyl alcohol, acetonitrile, dimethyl formamide, dimethyl sulfoxide, and triethylene glycol) as recommended in Guidance Document No. 23 (OECD, 2019) should be used whilst solvents of unknown toxicity should not be used. It should be noted that in spite of the low toxicity for fish, dimethyl formamide and dimethyl sulfoxide should be avoided where possible on human health and safety grounds.**

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium

characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

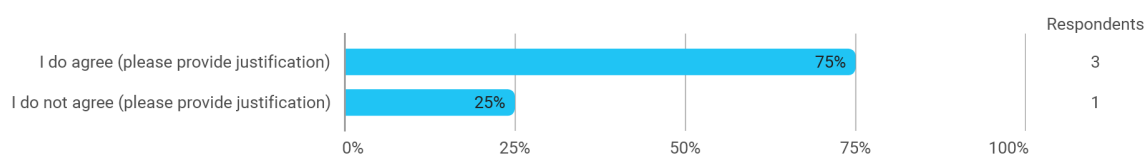
- Harmonization needed
- I'm okay with this. The wording needs to be consistent.

Disagreement justifications:

- DMF is a routinely used solvent in aquatic testing. It has been shown to be more suitable in most cases as e.g. acetone or ethanol. Generally, untested substances, as routinely tested in ecotox labs, have to be handled as if they were CMR compounds. Thus at least the safety measures in the labs should always be appropriate. Based on the fact that DMF was predominantly used by CRO's most of the historical data being available for possible effects of solvents, relate to DMF. The currently ongoing project investigating whether always a solvent control and a control is needed when a solvent is used in a study based their evaluations as well on a high number of studies which have been performed with DMF. The hope is, that this evaluation will allow to skip the negative control and by that reduce animal numbers in vertebrate tests. If we lose DMF that might have a negative impact on this decision.

5. Add Para. 18 from OECD TG 203.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

- Makes sense

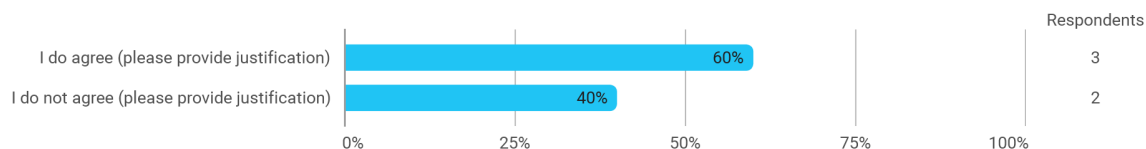
Disagreement justifications:

- The TG 203 para 18 calls for performing a test in parallel without pH adjustment if the pH of the stock solution must be adjusted to 6.0-8.5. That is impractical for a chronic fish study. Considering the fact that unnecessary vertebrate testing needs to be avoided this should not be considered for the FELS test. The pH adjustment in the range of 6 to 8.5 is allowed due to historical findings which demonstrated that this is an acceptable pH range for the used fish species. Thus, in case a pH adjustment is needed there is already sufficient information to

consider this and the use of additional fish to prove it again and again is not needed and should clearly be avoided.

6. Frequency of analytical determinations and measurements (Para. 35-40): align with Para 24-25 of OECD TG 203.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

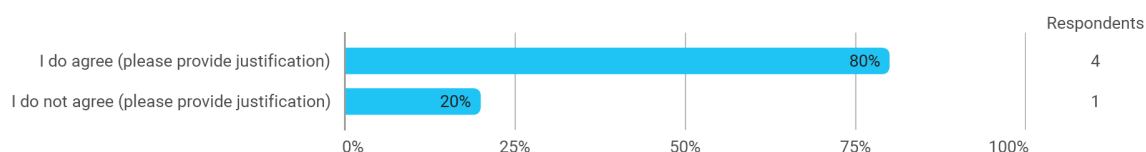
- Harmonization needed

Disagreement justifications:

- Disagree to align with paragraph 25 about the frequency of measurement of pH, O₂, salinity, etc. It is not necessary to measure them daily in all test vessels. Therefore, the para should stay unchanged. With respect to pH, O₂, salinity etc. The currently presented information is sufficient in most cases. Therefore, the adaptation of this para is not urgently needed. Agree to align with OECD 203 for para 24 regarding chemical analysis frequency.
- Two completely different studies with completely different sampling schedules.

7. Treatment of results: Add the following text from Para. 31 of OECD TG 203: It is recommended that results should be calculated using the measured concentrations of the test chemical. If the deviation from the nominal concentrations is smaller than 20%, results may also be based on the nominal concentrations. It should be noted that it is often useful to have both measured and nominal effect concentrations quoted, see Guidance Document No. 23 (OECD, 2019).

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

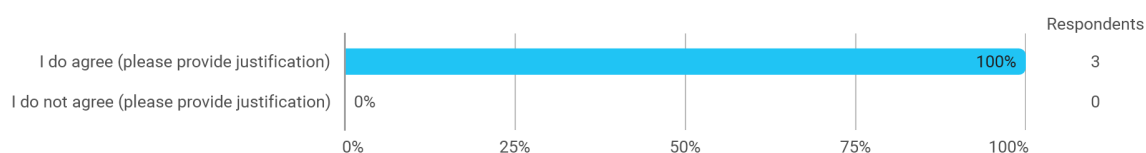
- Update needed
- Except for the last sentence, which I disagree adding to the TG.

Disagreement justifications:

- Disagree. The information given in para 24 is sufficient. Different regulatory frameworks use different approaches for the calculations. Harmonization would be appreciated but might not be the right way to do it via the OECD technical guidelines.

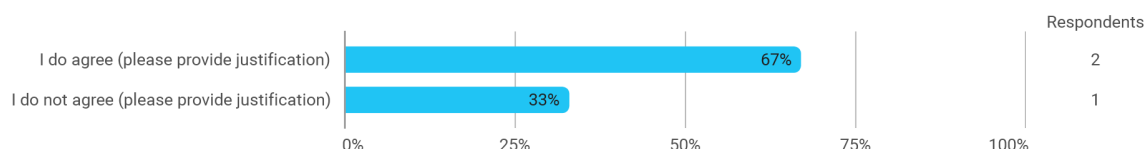
8. Annex 3 to be aligned with the Annex 3 from OECD TG 203 or at least include consistent information related to the content in particulate matter, TOC and COD in the description what makes an acceptable dilution/test water.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



9. para 2 - The sentence 'Alternatively, the data may be compared with control values ...' should say: 'In addition, the data should be compared with control values in order to determine ...'

Rational: According to the data requirements for active substances and product for plant protection EU 283/2013 and EU 284/2013, both, EC_x (ie EC₁₀ and EC₂₀) and NOECs need to be calculated and reported for long term/ chronic risk assessment.

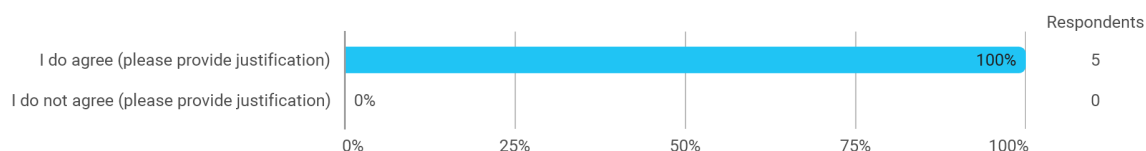


Disagreement justifications:

- This is an either or. That is if regression, controls are not compared. If trying to determine a LOEC, then controls need to be compared.

10. para 42 - Add wording from TG 210, paragraph 28: "Where deformities and associated abnormal behaviour are considered so severe that there is considerable suffering to the organism, and it has reached a point beyond which it will not recover, it may be removed from the test. Such animals should be euthanised and treated as mortalities for subsequent data analysis."

Rational: For reasons of consistency and animal welfare, instructions on removing animals from tests due to severe suffering should be harmonised.



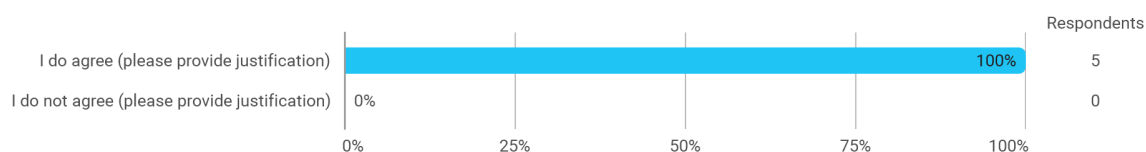
Agreement justifications:

- This addition is warranted for animal welfare reasons.
- update needed
- For the sake of consistency.

11. para 42 and 54 - Guideline OECD 215 requires the reporting of abnormal behavioural signs, but entirely lacks guidance on when and how to include such sublethal clinical observations (e.g., abnormal behaviour) in the long-term endpoint estimation. Can this be clarified, and guidance provided on how to include abnormal behaviour in the interpretation of the results? Does the NOEC / EC10 have to 'cover' certain abnormalities? The guidelines OECD 203 and 210 include somewhat more guidance aiming to extrapolate sublethal clinical signs to moribundity (which also requires updating, see comments made on these guidelines).

Rational: At the PRAPeR for difenoconazole such an issue is being discussed.

Sometimes the reported sublethal clinical signs suggest impending death and must thus be considered an acute effect, but are then not taken into account as there is not clear guidance on how and when to take them into account in the endpoint calculation.

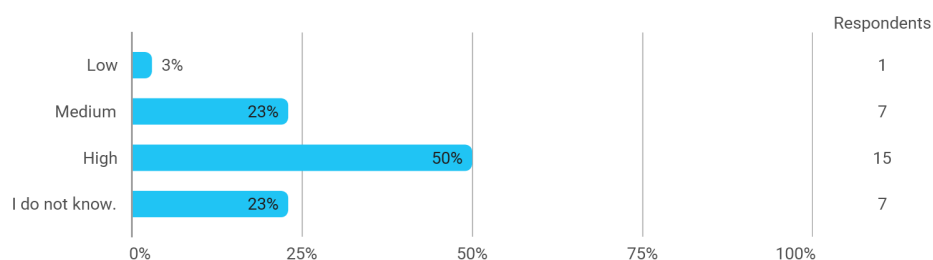


Agreement justifications:

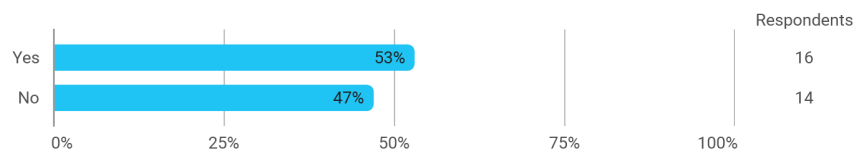
- For more specific information on the interpretation and application of clinical signs, further research and input from industry on their experience is needed.
- This would be helpful to the reader
- Clarity is always helpful

C.5 Test No. 229: Fish Short Term Reproduction Assay

Relevance of revision



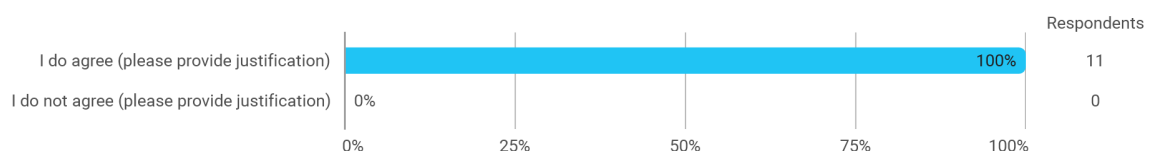
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. paras 24, 25 - The MTC should better defined and language aligned across all the relevant aquatic endocrine TGs.

Rational: There is significant confusion as to the methodology to set an MTC. All the relevant TGs use different language so there is no clear consensus. This leads to regulatory uncertainty with may mean tests are performed at too high or low test levels. This may trigger unnecessary requests for repeat or higher tier tests.



Agreement justifications:

- Better defined, but MTC must consider TG exposure duration
- ▶ or refer to literature, but the concept of MTC should it self be further discussed and researched given the non monotonic response of endocrine sensitive endpoints
- ▶ MTC guidance ist needed to provide better setting of concentration range well below systemic toxicity
- ▶ more clarity would be helpful
- ▶ This is an issue across all OECD fish and amphibian TGs and I would strongly support efforts to harmonise concentration setting., We are currently conducting a data analysis project to support this.
- ▶ Additional information for establishing the MTC would be helpful and a consideration of responses other than mortality should be considered for describing a MTC. There may be a need to specifically consider the definition of a MTC in assays that are designed to assess potential for endocrine activity and adverse effects mediated through endocrine activity.
- ▶ a better definition of the MTC could harmonize evaluation
- ▶ Harmonization needed
- ▶ strongly agree this the major uncertainty when designing these studies

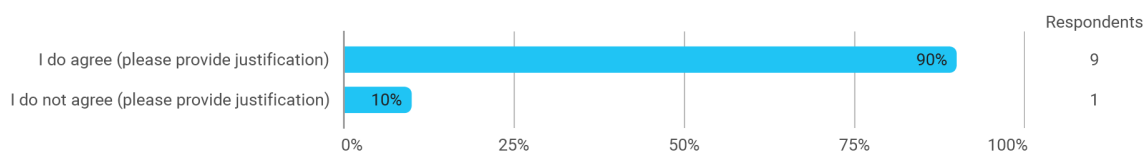
2. para 49 - A clear description of how statistical analysis for each parameter should be performed. Each parameter should have a defined analysis strategy i.e. one-tailed (lower or greater) or two-tailed. A table format is suggested.

Rational: The current TG demonstrates no clear strategy for the analysis of each endpoint. This has led to different CROs performing different analysis for endpoints,

and a general lack of consistency. A clear description of whether a one-tailed or two-tailed analysis should be denoted against each endpoint.

It is currently not clear whether two-tailed analysis should be used when an endpoint has the biological potential to change in either direction, OR, based on only the relevant change for conclusion of ED. For example, plasma VTG can biologically both increase and decrease in males (although basal in control males), however a decrease in VTG in males would not be relevant for assessment of endocrine potential.

Therefore should this parameter be analysed one-tailed (greater) or two-tailed?



Agreement justifications:

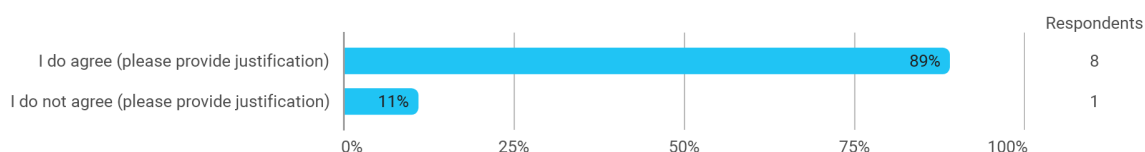
- ▶ this could be helpful.
- ▶ The use of different statistical approaches may lead to differential assessments and some studies more stringent than others
- ▶ Further advise is needed to provide an efficient statistical evaluation of the data
- ▶ more clarity would be helpful
- ▶ referring to OECD 54
- ▶ This would be helpful to the user
- ▶ Agree can make a big difference to outcome so should be clearly specified

Disagreement justifications:

- ▶ Getting overly-prescriptive on how the stats should be handled may result in loss of flexibility for special cases or in the event of new information that changes how we approach the analysis.

3. **para 38 - Add wording from TG 210, paragraph 28: “Where deformities and associated abnormal behaviour are considered so severe that there is considerable suffering to the organism, and it has reached a point beyond which it will not recover, it may be removed from the test. Such animals should be euthanised and treated as mortalities for subsequent data analysis.”**

Rational: For reasons of consistency and animal welfare, instructions on removing animals from tests due to severe suffering should be harmonised.



Agreement justifications:

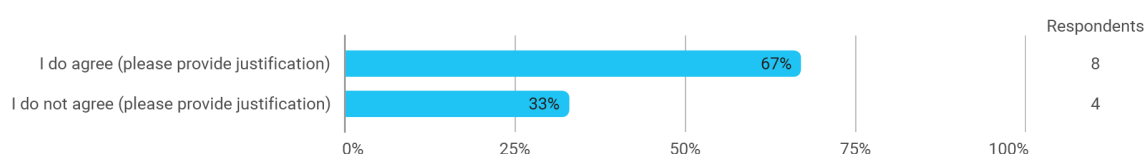
- ▶ Animal well fare standards must be met. Such cases are not expected in a well designed study, and mortality should not be related to test item so no consequences to the test results are expected
- ▶ important to include sentence to be in alignment with local animal welfare regulations
- ▶ This addition is warranted for animal welfare reasons.
- ▶ Align with other TGs
- ▶ However, there should be no sublethal effects observed during this test.
- ▶ Update needed
- ▶ Any opportunity to strengthen animal welfare considerations should be taken

Disagreement justifications:

- ▶ While efforts to eliminate animal suffering are supported, it is suggested is to wait until there is a clear understanding of fish morbidity. Since there may be variability in how fish suffering is assessed among different laboratories and this may contribute to additional variability in study conduct. The definition of fish morbidity is currently under investigation through data collected in the conduct of the fish acute toxicity test (OECD 203).

4. para 24 - Suggestion to add to paragraph 24: If a range finder is employed and a solvent carrier is required, it is acceptable to use only one control group i.e. a solvent control. The addition of a clean water control is not necessary.

Rational: In the interest of animal use, where range finding studies are required, there should not be a requirement to run two control groups (dilution water and solvent control) where solvent is used. This should be made clear in the TG.



Agreement justifications:

- ▶ Yes with the intention of reducing animal use and because many non toxic solvents are available are
- ▶ This addition is warranted for practical and animal welfare reasons.
- ▶ 3Rs, reducing number of fish tested.
- ▶ removing the control if a solvent control is used can reduce the number of fish used in a range finder and is in accordance to the 3r principles
- ▶ However it also depends on the type of solvent used (will known or not).

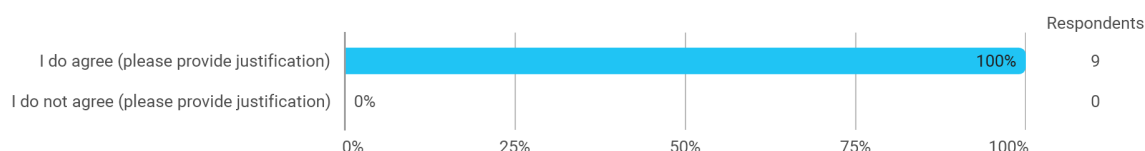
Disagreement justifications:

- ▶ I do not agree in not using a negative control while using a solvent control, since effects of the solvent could not be detected at this point and probably blur effects of the test item.

- ▶ I do not agree
- ▶ I agree if we're only talking about the range finder. The definitive test must include a clean water control, as you never know whether your solvent can be contaminated by something or not.
- ▶ Although this statement supports the idea that the comparison should be to the solvent control (which is supported), this change would result in lack of harmonization with other regions. Maybe the wording could be “dilution water control is optional when solvent control is used”. Also, having double the number of controls (in case of pooling) provides additional information on the characterization of the population response.

5. para 24 - The recommendation is to avoid any systemic toxicity in definitive studies with three test concentrations. Clearer text is needed to highlight this.

Rational: Some regulators are expecting to see some systemic toxicity from the highest test concentrations, based on current text in TG. However, testing to the MTC can mean that the highest test concentration is often lost to systemic toxicity, and therefore not available for the interpretation of ED potential. Given only three concentrations are used in the test, this leaves only two test concentration for interpretation of ED potential. Assessing ED potential from only two test concentrations can be problematic. A clearer definition of testing to the MTC should be outlined in the TG. The aim should be to avoid any systemic toxicity in definitive studies. Existing data and/or range finders (where chronic species data is not available) should be used to position the definitive test concentrations to avoid systemic toxicity effects.

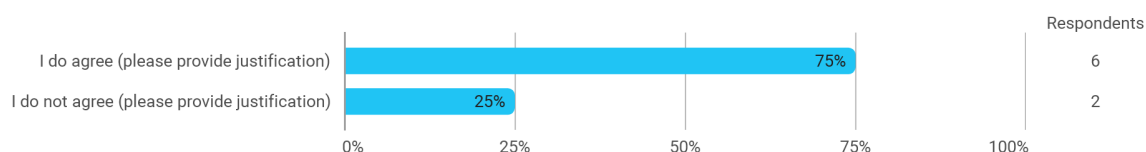


Agreement justifications:

- ▶ This may be clarified, yes
- ▶ see suggestion 1
- ▶ This addition is warranted to prevent systemic toxicity from confounding endpoints and for animal welfare reasons.
- ▶ more clarity would be helpful
- ▶ Additional language would be helpful to clarify that systemic toxicity is not desired in this study. However, there is still a need to maintain the criteria that the test is still valid if one of the test levels is lost due to systemic toxicity. It is not always the case that any signs of systemic toxicity are avoided in this study.
- ▶ please also see comments in point 3
- ▶ clarification would be helpful to the user
- ▶ associated with poor MTC definition across relevant TGs

6. para 53 - Recommendation to add sentence to paragraph 53 to include acceptable levels of natural variability in VTG data.

Rational: Experience shows that VTG can be highly variable, however there is no guidance in the TG on natural variability and acceptability of data.



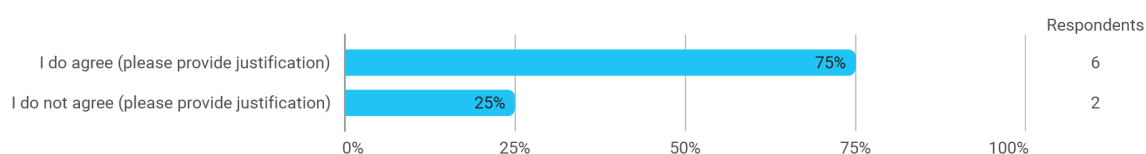
Agreement justifications:

- sometimes hard to evaluate VTG results due to very high natural variability (effects may not be statistically significant due to high variability)
- more clarity would be helpful
- However rather than "acceptable" level (it might depend on the test assessor), information could be given on the natural variability inherent to the endpoint
- Agree but should be handled carefully and not be too prescriptive

Disagreement justifications:

- that some discussion is needed concerning this endpoint but acceptable levels of natural variation is unclear. Discussion is required and the use of historical control data should be a tool to consider
- It is not clear what the natural variability of VTG levels are, especially since this parameter may vary based on lab and analytical methodology used to quantify VTG. The test guidelines should be flexible with regard to these considerations

7. para 13 - Language offering more flexibility around the test performance criteria should be included.



Agreement justifications:

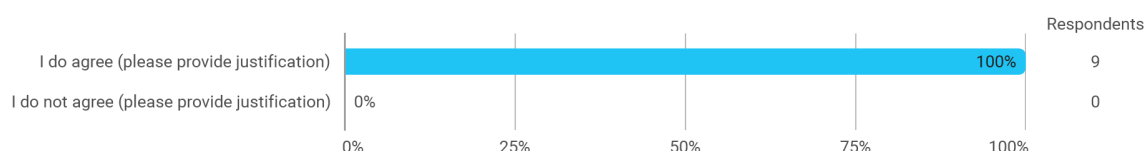
- Some validity criteria (namely analytics) should be considered performance criteria
- Agree more flexibility should be given for fish studies to avoid unnecessary repeats.
- More flexibility is good in this regard. For example. more flexibility around slight dissolved oxygen and/or temperature deviations should be included. In addition, especially in the case of test substances that are difficult to test in aquatic systems, there should be greater acceptance if concentrations of the test material cannot be maintained within 20% of mean measured.
- an additional point should be added – if one of the VC are not met, then assessment should be added.

- clarity is required as the TG does not have 'hard' validity criteria

Disagreement justifications:

- I do not agree
- Test criteria should be met. I am not sure how flexible we can be with that. By being too flexible, we might end up in a situation where invalid tests are submitted.

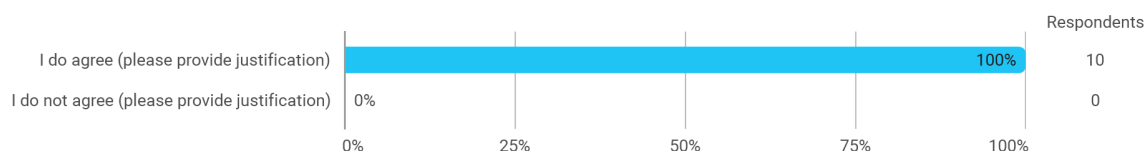
8. para 24 - Test concentration setting guidance (The MTC) needs to be improved. It should be made clearer that mortality is not desirable. Suggest that MTC guidance is aligned amongst all relevant test guidelines.



Agreement justifications:

- Better defined, but MTC must consider TG exposure duration
- same as suggestion 1
- see suggestion 1
- same as for suggestion 1
- Additional information for establishing the MTC would be helpful and a consideration of responses other than mortality should be considered for describing a MTC. There may be a need to specifically consider the definition of a MTC in assays that are designed to assess potential for endocrine activity and adverse effects mediated through endocrine activity.
- see point 5
- Yes, a better definition and harmonization across the TG is definitely needed
- strongly agree see comment above

9. Annex 2/3 - validity criterion: test concentrations within measured concentrations $\pm 20\%$ should be further clarified, particularly in the case of unstable or volatile substances of which the concentration is not technically possible to maintain stable in a flow through system



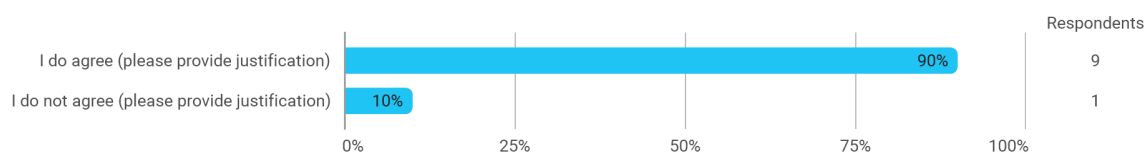
Agreement justifications:

- I agree. if unstable substances are used the test can be invalid directly at the test start, even though this would be the best application way of the test item. If the 20% range would be a recommendation, not a validity criteria this would be more useful. in cases where the 20%

range is not fitted, a geomean or time weighted average could be reported and this could be stated in the guideline.

- ▶ agree
- ▶ A major problem of this TG, unnecessarily invalidating many studies , reconsider as validity criteria, performance criteria might be better. Criterion could be made more flexible and adjustable to different types of substances
- ▶ I do agree
- ▶ Having analytical as validity is very troublesome when dealing with UVCBs and insoluble compounds.
- ▶ Perhaps some wording could be added like “...evidence should be available to demonstrate that the concentrations of the test substance in solution have been satisfactorily maintained within $\pm 20\%$ of the mean measured values with consideration given to difficult to test substances...”
- ▶ especially important for vertebrate studies – when does a deviation from VC will invalidate the study? This should be clarified
- ▶ please change the chemical analysis criterion to a performance criterion, because in some cases chemical analysis within $\pm 20\%$ is impossible
- ▶ This could be helpful to the user

10. para 21 - Broaden the recommended age range of fish

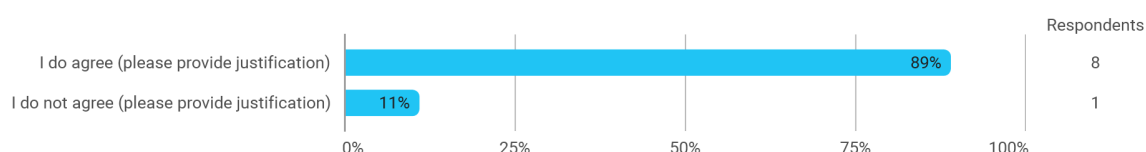


Agreement justifications:

- ▶ The fish age used at the moment leads to an increased breeding of fish to get animals in the correct age and as the correct size and weight. therefore to small fish can not be used for the test but can also not be used later because of the age range. This is not in accordance to the 3R principles and reducing number of animals bred and used for tests.
- ▶ strongly agree. The age range should reflect the fish used in validation studies
- ▶ Yes, such a narrow age range leads to discarding a large part of the fish production, Consider at least the interval suggested in OCSP890.1350 (EPA)
- ▶ actively spawning fish more important than age
- ▶ we have found that fish at the lower age range are not obviously male or female.
- ▶ This could be helpful to increase flexibility and reduce animal use. Some input from contract labs/EPA ORD would be helpful here regarding the range for each species. Demonstration of adequate reproduction is still needed.
- ▶ but also to include the assessments of spawning and also should include the broadening of weight range.

- broaden the recommended age of the fish decreases the fish disposed because of the wrong age and is in accordance with the 3r principles

11. Annex 2 / paras 18 and 23 - Aeration should be allowed since test start. Oxygen dropping below 60% renders the test invalid, however aeration is only permitted if oxygen drops below 60%



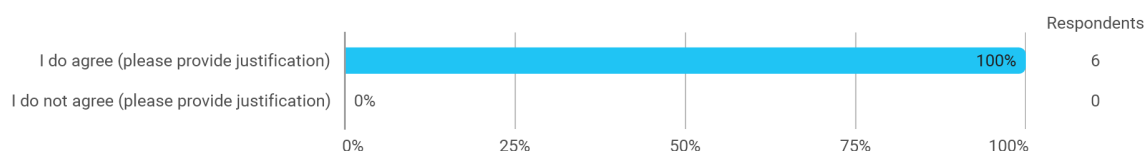
Agreement justifications:

- aeration should be applied as soon as it is needed or as soon as the scientist thinks it should be used in advance for good conditions for the fish
- Contradicting validity criteria and permitted actions
- check of oxygen concentration once a week without additional aeration might lead to non-compliance with validity criteria
- oxygen should be added to methods to avoid validity criteria issues
- it's risky not adding aeration to a chronic exposure
- it is totally reasonable to add aeration from the beginning of the test because this does not influence the fish in a bad way and decreases the risk of deviations to the guideline
- update needed to be coherent

Disagreement justifications:

- Aeration may not be appropriate at test start if the test material is volatile. Thus, a rewording to allow aeration if there is a concern or a trend toward lowering DO concentrations would be helpful. This may also be addressed by revising test acceptance criteria and allowing for more flexibility.

12. Annex 2/4 - additional information on Wide spectrum light (several types possible fluorescent bulbs, LED, etc.)

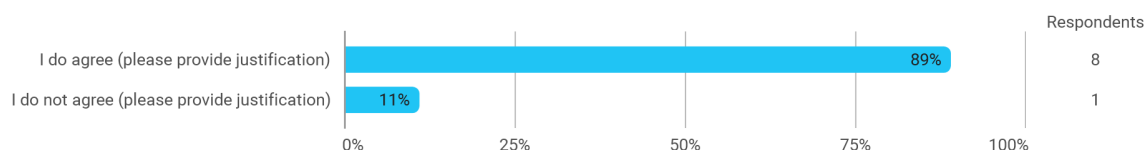


Agreement justifications:

- Yes, but of minor concern
- more flexibility for laboratory set up
- The option here would be to describe what wide spectrum is characterized as and let the laboratories choose the type of light fixture.
- some examples can be given of acceptable lighting spectrum

- more detailed information is helpful

13. para 21 - addition of wider range of age of fish in combination with egg production



Agreement justifications:

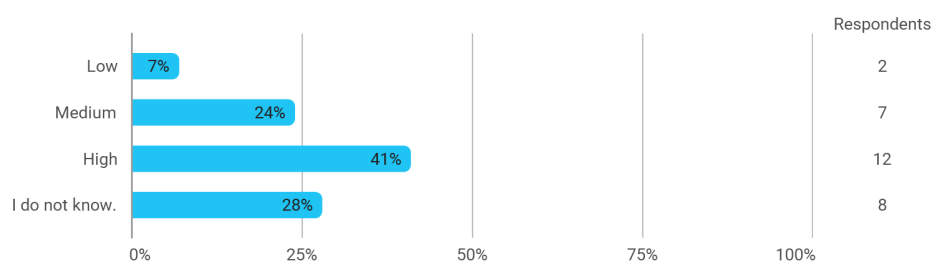
- strongly agree. The age range should reflect the fish used in validation studies
- But requires clarification
- actively spawning fish more important than age
- practically this would be very helpful and reduce the need to breed specifically for this study. therefore timelines would be reduced.
- This could be helpful to increase flexibility and reduce animal use. Some input from contract labs/EPA ORD would be helpful here regarding the range for each species. Demonstration of adequate reproduction is still needed.
- see point 10
- see suggestion 10

Disagreement justifications:

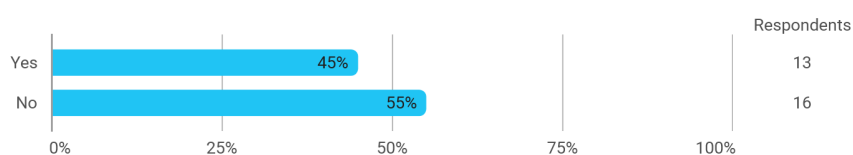
- I am not strictly against this, however I don't understand what would be the added value of this? I think this should be discussed.

C.6 Test No. 230: 21-day Fish Assay

Relevance of revision

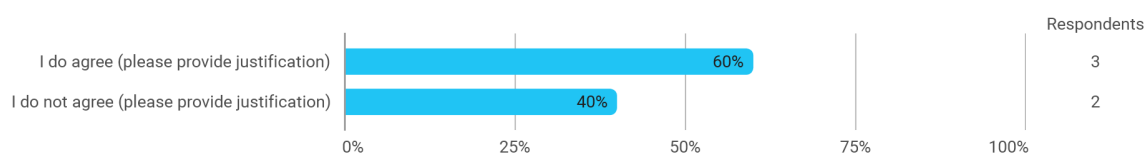


Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. para 12 - Language offering more flexibility around the test performance criteria should be included.



Agreement justifications:

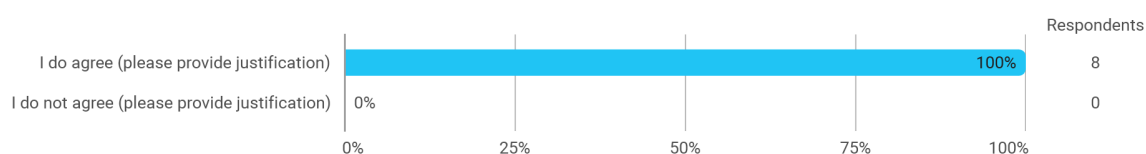
- ▶ More flexibility around the test performance is generally viewed as good in order to potentially reduce animal use via eliminating unnecessary study repeats. It is unclear what is specifically meant here, however.
- ▶ an additional point should be added – if one of the VC are not met, then assessment should be added.

Disagreement justifications:

- ▶ I do not agree
- ▶ Validity criteria should be met. By being too flexible, we might end up with invalid test being submitted

2. paras 23, 24 - The MTC should be better defined and be consistent across relevant aquatic endocrine TGs

Rational: To avoid regulatory uncertainty and unnecessary repeat or higher tier tests.

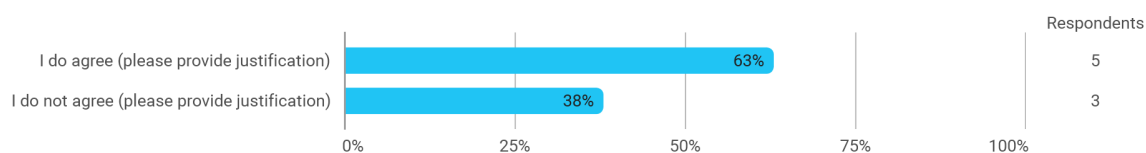


Agreement justifications:

- ▶ Better defined, but MTC must consider TG exposure duration
 - ▶ more guidance needed to determine MTC
 - ▶ see comments on TG 229
 - ▶ Additional information for establishing the MTC would be helpful and a consideration of responses other than mortality should be considered for describing a MTC. There may be a need to specifically consider the definition of a MTC in assays that are designed to assess potential for endocrine activity and adverse effects mediated through endocrine activity.
 - ▶ more detailed information on the MTC is helpful
 - ▶ update needed
- 3. para 23 - Recommendation to add to paragraph 23: Where a range finding study is employed and a solvent carrier required, only one control group is required. A solvent**

control group can be employed, without the need to use an additional clean water control.

Rational: In the interest of animal use, where range finding studies are required with a solvent carrier, only a solvent control group should be required. An additional clean water control is not required.



Agreement justifications:

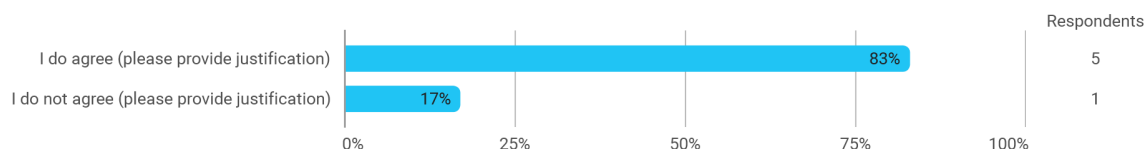
- This addition is warranted for practical and animal welfare reasons.
- removing a control group in the rangefinder doesn't affect the outcome but reduces the number of fish used and is in accordance to the 3r principles
- Update needed

Disagreement justifications:

- I agree for the RF only. For the definitive test, a clean water control must be included as you never know whether your solvent is contaminated by endocrine active chemicals or not.
- Although this statement supports the idea that the comparison should be to the solvent control (which is supported), this change would result in lack of harmonization with other regions. Maybe the wording could be "dilution water control is optional when solvent control is used". Also, having double the number of controls (in case of pooling) provides additional information on the characterization of the population response.

4. para 23 - Recommendation if a range finding study is employed, guidance on how many fish should be used per treatment should be added.

Rational: TG should offer guidance on animal numbers and replication (if relevant) for range finders. This will help CROs justify animal use and apply for relevant permits, and standardise the numbers used across CROs. Current restrictions on animal usage vary between countries, and therefore there is not currently a standard approach to running range finders.

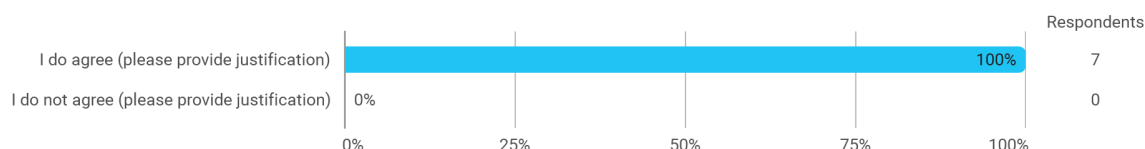


Agreement justifications:

- guidance would be appreciated, right now number of fish are chosen in concordance with fish number per replicate in 21-Day fish assay
- This addition is warranted for practical and animal welfare reasons. It requires input from those experienced with conducting the test.
- this would be helpful tot he user

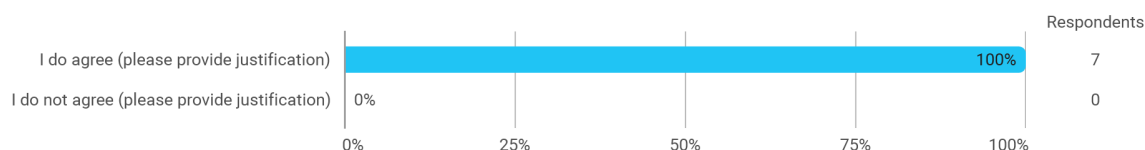
Disagreement justifications:

- ▶ Prefer not to be overly prescriptive on the conduct of range - finding studies. The needs will differ based on the test material and the available amount of toxicity data.
- 5. para 23 - Test concentration setting guidance (The MTC) needs to be improved. It should be made clearer that mortality is not desirable. Suggest that MTC guidance is aligned amongst all relevant test guidelines.**



Agreement justifications:

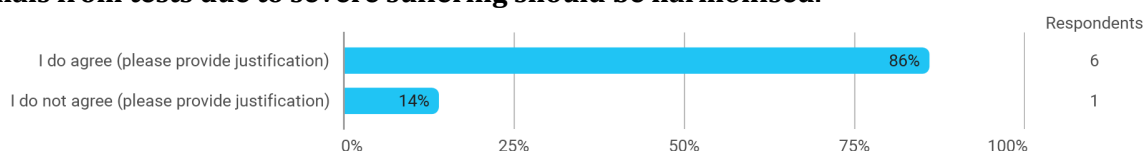
- ▶ Better defined, but MTC must consider TG exposure duration
 - ▶ more guidance needed to determine MTC
 - ▶ This addition is warranted to prevent systemic toxicity/mortality from confounding endpoints and for animal welfare reasons.
 - ▶ Additional information for establishing the MTC would be helpful and a consideration of responses other than mortality should be considered for describing a MTC. There may be a need to specifically consider the definition of a MTC in assays that are designed to assess potential for endocrine activity and adverse effects mediated through endocrine activity.
 - ▶ update definitely needed
- 6. para 24 - Recommendation to include a clear definition of testing to the MTC - systemic toxicity should be avoided if possible in definitive studies, due to the impact on interpretation of endocrine potential of a test substance. Existing data and/or range finders (where required) should be used to denote the MTC, and definitive studies should be testing just below MTC.**
- Rational:** Testing to the MTC can mean that the highest test concentration is often lost to systemic toxicity, and is therefore not available for the interpretation of potential ED effects. Given the test is conducted with three test concentrations, this leaves only two concentrations for interpretation of ED potential, which can be difficult (although acceptable). There are also animal welfare implications when systemic toxicity is present in a study. The recommendation is to avoid systemic toxicity in definitive studies where possible, using existing chronic species data where available to position the MTC.



Agreement justifications:

- ▶ Better defined, but MTC must consider TG exposure duration

- ▶ more guidance needed to determine MTC
 - ▶ This addition is warranted to prevent systemic toxicity from confounding endpoints and for animal welfare reasons.
 - ▶ Some additional guidance around the MTC would be good so that we are not expected to kill fish at the top level, but we need to maintain the criteria that the test is still valid if one of the test levels is lost due to systemic toxicity. It is not always possible to avoid any signs of systemic toxicity in the study design and there is pressure to test at high concentrations to make sure that ED is not missed by testing too low, so keeping the ability to drop the top dose and still have a valid study is helpful.
 - ▶ Update and clarification needed
7. **para 37 - Add wording from TG 210, paragraph 28: “Where deformities and associated abnormal behaviour are considered so severe that there is considerable suffering to the organism, and it has reached a point beyond which it will not recover, it may be removed from the test. Such animals should be euthanised and treated as mortalities for subsequent data analysis.”**
Rational: For reasons of consistency and animal welfare, instructions on removing animals from tests due to severe suffering should be harmonised.



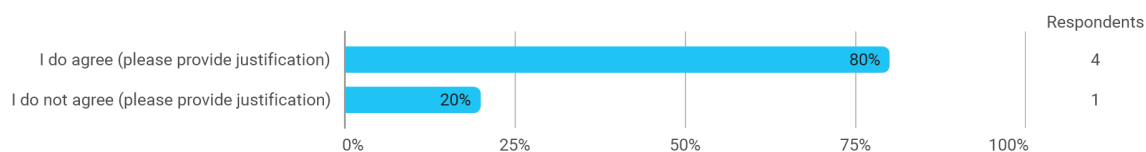
Agreement justifications:

- ▶ sentence needed to allow alignment with local animal welfare regulations
- ▶ This addition is warranted for animal welfare reasons.
- ▶ in line with the 3Rs principles
- ▶ there should be no sublethal effects observed during this test.
- ▶ Update needed

Disagreement justifications:

- ▶ While efforts to eliminate animal suffering are supported, it is suggested is to wait until there is a clear understanding of fish morbidity. Since there may be variability in how fish suffering is assessed among different laboratories and this may contribute to additional variability in study conduct. The definition of fish morbidity is currently under investigation through data collected in the conduct of the fish acute toxicity test (OECD 203).
8. **para 49 - Recommendation is to include guidance in paragraph 49 on how the analysis of each parameter should be done i.e. one-tailed (greater or lower) or two-tailed. A table format with the recommendation for analysis (one or two-tailed) of each parameter is suggested for inclusion.**
Rational: Current TG offer no guidance on how the analysis of each endpoint should be done. This has led to different CROs using different analyses for the endpoints, leading to a lack of consistency. It also needs to be clearer in the TG what the analysis should

be based on - i.e. one or two sided analysis based on biologically possible changes, OR, based on the relevance of the change for the assessment of endocrine potential. For example, plasma VTG in males can both increase and decrease biologically (although basal in control males), however only an increase in plasma VTG is relevant for assessing endocrine potential on this parameter in males. Therefore, would the recommendation be for a one-tailed (greater) or a two-tailed analysis?



Agreement justifications:

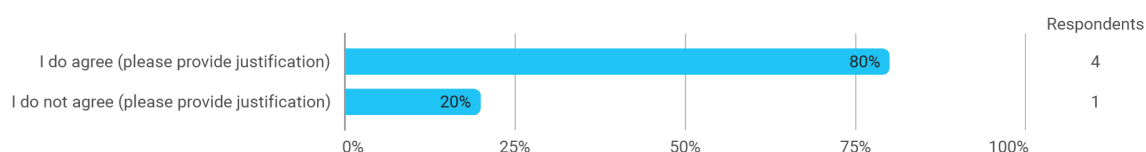
- ▶ more precise decision tree for statistical analysis needed
- ▶ This would be helpful to the user

Disagreement justifications:

- ▶ Getting overly-prescriptive on how the stats should be handled may result in loss of flexibility for special cases or in the event of new information that changes how to approach the analysis.

9. para 53 - Recommendation to add a sentence to paragraph 53 to outline acceptable natural variability in VTG data.

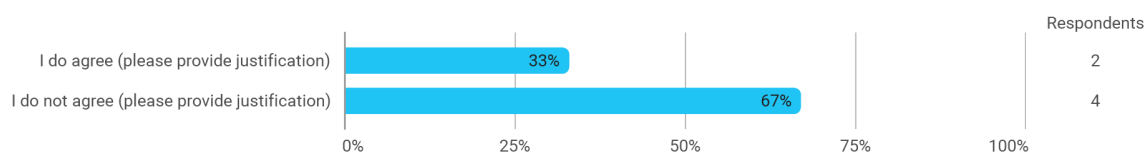
Rational: VTG data has demonstrated to be highly variable, however there is no guidance in the TG of acceptable levels of variability. This is important to add, in order to give regulators and applicants clear view of the level of acceptable natural variability in this endpoint.



Agreement justifications:

- ▶ recommendation for high natural variability needed to allow better comparison with treatment groups (statistical evaluation)
- ▶ Not sure if there is a consensus on what "acceptable" would be but some information could be given on the natural variability of the endpoint
- ▶ It is not clear what the natural variability of VTG levels are, especially since this parameter may vary based on lab and analytical methodology used to quantify VTG. The test guidelines should be flexible with regard to these considerations.

10. The methodology as described in the current guideline is not suitable to test for effects of virtually insoluble particulate materials (e.g. nano-forms of substances, polymers). Adaption of the guidelines methodology with latest scientific findings is indicated so that valid tests with particulate materials can be performed.



Agreement justifications:

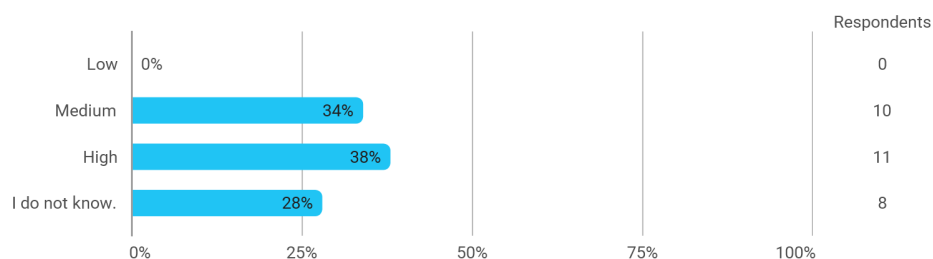
- ▶ guideline should be up to date with scientific standards

Disagreement justifications

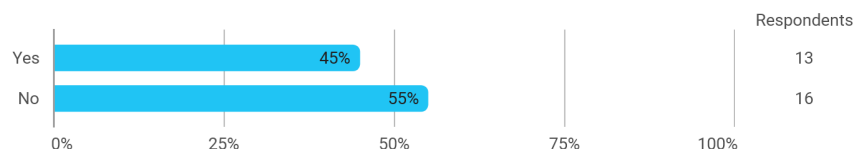
- ▶ See OECD GD 23 and OECD GD for aquatic toxicity testing of manufactured nanomaterials
- ▶ Some general wording to allow for testing a diversity of materials could be included (i.e. for testing nanomaterials or polymers above water solubility limit etc.); however detailed instructions on testing difficult substances should not be addressed in individual toxicity testing guidelines, but should be preferentially addressed in guidance documents such as OECD 23.
- ▶ Methodology is not available yet. This should be a separate guidance
- ▶ Not sure this could be done so easily; wouldn't require additional validation work?

C.7 Test No. 234: Fish Sexual Development Test

Relevance of revision



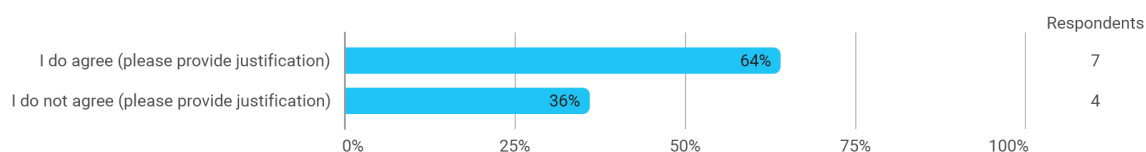
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. It is suggested to add thyroid-specific endpoints to the guideline. Specifically: thyroid hormone measurements, thyroid histopathology, eye histopathology and assessment of swimbladder inflation.

Rational: The above listed endpoints are well investigated in fish and are responsive to thyroid disruption. There is no scientific reason why the T-modality should not be assessed in fish, especially in developmental exposure experiments according to TGs 234, 210 or 236. Multiple early developmental processes in fish (incl. metamorphosis from embryo to juvenile) are regulated by thyroid hormones and can be disrupted by different endocrine disruptors.



Agreement justifications:

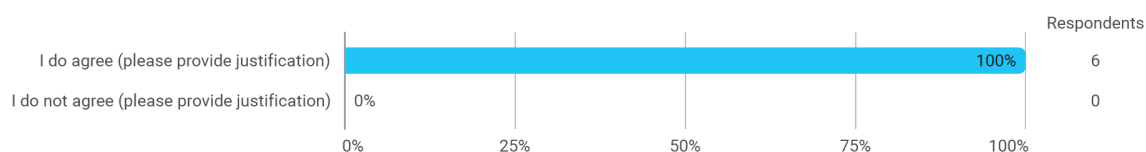
- ▶ it is important to gain more information concerning endocrine activity of substance without increase in fish numbers would be useful
- ▶ This is decreasing the information needs for thyroid disruption
- ▶ this would need to be fully validated.
- ▶ with a view to eliminating separate amphibian studies, providing it does not add a welfare burden to the fish test
- ▶ Update needed when enough background knowledge will be gained on these endpoints
- ▶ But these need to validated before inclusion - the especially the animals that might be required for T hormone measurements (see Wheeler et al 2021 Env Int 146: 106287)

Disagreement justifications:

- ▶ I do not agree to add such huge modifications to the TG, unless a huge literature review is performed, showing that the science is sufficiently mature to add such endpoints.
- ▶ A number of questions need to be discussed before decisions on the inclusion of thyroid-specific endpoints can be made: Would this increase the number of animals in the test? How conclusive can these endpoints be? Would a definite result on thyroid-related activity be possible or would further testing be required?
- ▶ There should be a clear understanding of the specificity of the endpoint for indicating thyroid pathway perturbation in addition to a formal validation of the measurement before adding this to the test guideline.
- ▶ until this protocol is established, this should not be included.

2. Include a section on initial considerations in line with the latest update of OECD TG 203 (18 June 2019).

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



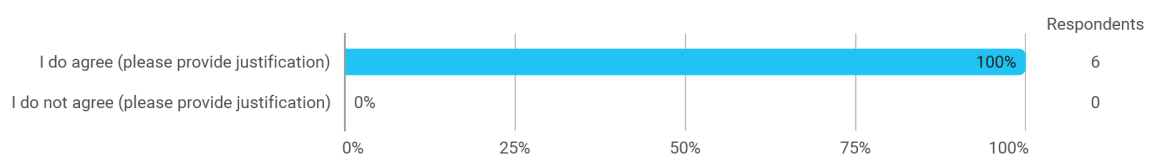
Agreement justifications:

- ▶ to avoid problems with test item, initial considerations concerning the test item should be included in guideline

- ▶ align with 203
- ▶ agree
- ▶ harmonization needed

3. para. 17: add the following text from OECD TG 203: When testing difficult test chemicals, Guidance Document No. 23 (OECD, 2019) should be consulted, and the study design modified appropriately. For volatile and other difficult test chemicals, further specific measures should be taken (OECD, 2018). Any silicone materials that were in contact with the test solution(s) should be preferably discarded and not reused for subsequent tests with different test chemicals.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.

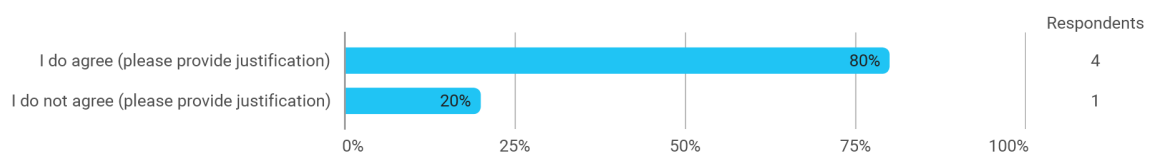


Agreement justifications:

- ▶ more precise information
- ▶ align with 203
- ▶ agree
- ▶ update and harmonization needed

4. para 22: align the wording with OECD TG 203. OECD TG 211 states that measuring TOC/DOC is only recommended while TG 203 states it should be measured. It is proposed to use the following wording from OECD TG 203: If natural water is used, DOC or TOC and nitrate-content (NO₃) should be measured once prior to the test.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.

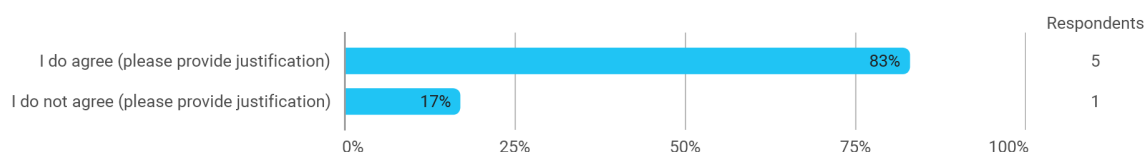


Agreement justifications:

- ▶ align with 203
- ▶ update and harmonization needed

Disagreement justifications:

- Labs are generally doing periodic tests of DOC/TOC and nitrate on a regular basis. These tests may not be needed before each test is conducted.
5. **para. 24: to be replaced by Para. 17 from OECD TG 203. The following text from Para. 23 of OECD TG 203 should also be included: Low toxicity solvents only (i.e. acetone, ethanol, methanol, tertiary-butyl alcohol, acetonitrile, dimethyl formamide, dimethyl sulfoxide, and triethylene glycol) as recommended in Guidance Document No. 23 (OECD, 2019) should be used whilst solvents of unknown toxicity should not be used. It should be noted that in spite of the low toxicity for fish, dimethyl formamide and dimethyl sulfoxide should be avoided where possible on human health and safety grounds.**
- Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.**

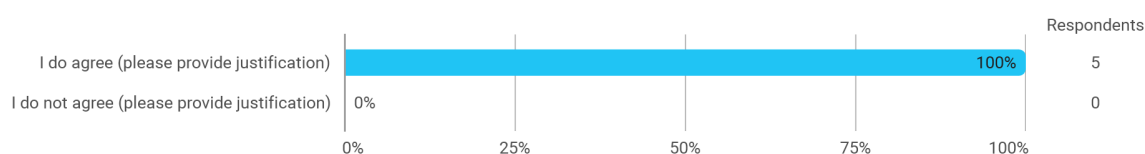


Agreement justifications:

- more precise wording needed concerning the use of a solvent
- align with 203
- update and harmonization needed

Disagreement justifications:

- DMF is a common solvent in ecotoxicity testing based on performance in assays with regard to low toxicity to aquatic organism and low biofilming in the studies. OECD test guidelines are international and should reflect a level of flexibility to accommodate the use of different solvents in different regions. It should be noted that protective controls for testing materials of unknown toxicity should be in place in laboratories such that solvent use would also be conducted in a safe manner
6. **Add Para. 18 from OECD TG 203.**
- Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.**

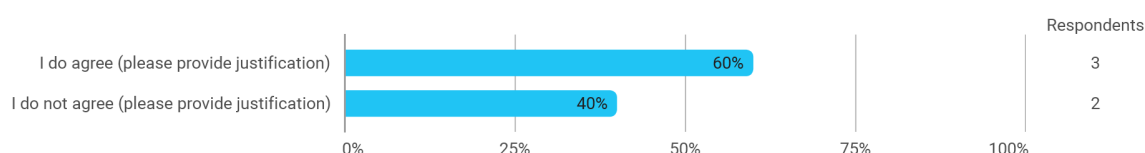


Agreement justifications:

- More precise guidance concerning substance handling needed
- align with 203
- update and harmonization needed

7. Frequency of analytical determinations and measurements (Para. 24-26): align with Para 24-25 of OECD TG 203.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

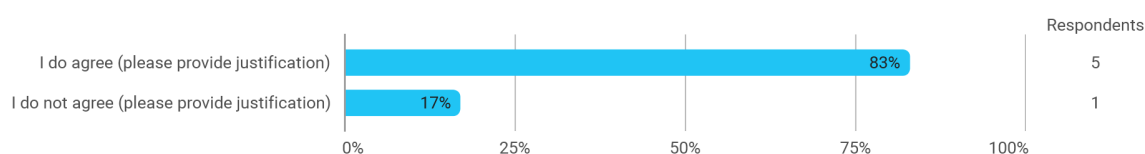
- More precise guidance concerning analytics needed
- align with 203

Disagreement justifications:

- If alignment of frequency of analytical and water quality measurements is needed, shouldn't it be with a longer duration study like the FSTRA or the ELS instead of a 4 day acute study? Also, paragraph 24 in OECD 203 refers to the LC50 which is not an endpoint goal in longer term studies.
- the existing paragraphs are appropriate for this study.

8. Treatment of results: add the following text from Para. 31 of OECD TG 203: It is recommended that results should be calculated using the measured concentrations of the test chemical. If the deviation from the nominal concentrations is smaller than 20%, results may also be based on the nominal concentrations. It should be noted that it is often useful to have both measured and nominal effect concentrations quoted, see Guidance Document No. 23 (OECD, 2019).

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

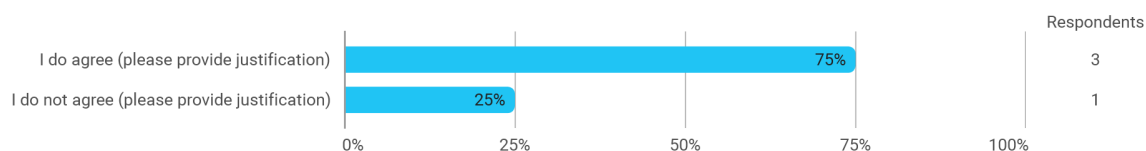
- ▶ more precise reporting on test concentration needed for better understanding o
- ▶ align with 203
- ▶ agree
- ▶ update needed

Disagreement justifications:

- ▶ Not applicable to this test design.

9. Align Annex 2 with the Annex 3 from OECD TG 203 or at least include consistent information related to the content in particulate matter, TOC and COD in the description what makes an acceptable dilution/test water.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



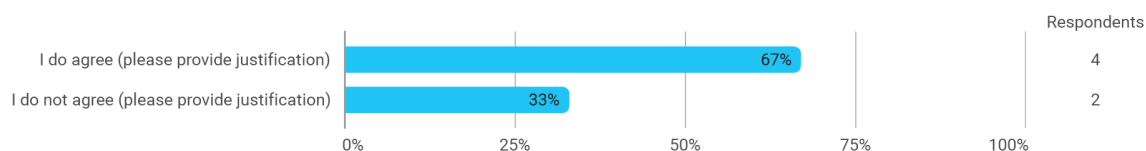
Agreement justifications:

- ▶ align with 203

Disagreement justifications:

- ▶ If alignment of acceptable water quality is needed, shouldn't it be with a longer duration study like the FSTRA or the ELS instead of a 4 day acute study?

10. para 1 - Please exchange 60 dph by until final maturation



Agreement justifications:

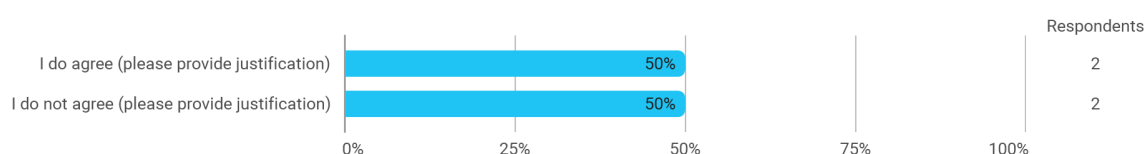
- ▶ reaching sexual maturity is dependent on various variables, therefore a change of wording is necessary
- ▶ This will ensure that the test animals have proceeded for the sufficient maturation for the ED assessment. Too young fish at the end of the test would make it more difficult to assess ED relevant parameters.
- ▶ in can be done either at 60 dph or final maturation depending on the spp used.

Disagreement justifications:

- Unclear what final maturation is referring to here. Does using this terminology introduce uncertainty and potential inconsistencies between laboratories? Suggest using more concrete timelines to ensure that the studies are conducted similarly at different labs.

11. para 16 - One of the validation criteria of the guideline refers to sex ratio (% males or females) in the control. Ranges 30-70% are reported for the three species. It should be considered to specify that the range of percentage refers to the overall fish in the control and not to only to the mature fish.

Rational: the current validation criterion is a bit ambiguous and should be clarified as it seems that in some of the recently performed studies with zebrafish the percentage of undifferentiated fish was high and the validity control was calculated only on mature fish. Is this how this validation criterion was intended? please consider clarifying this aspect.

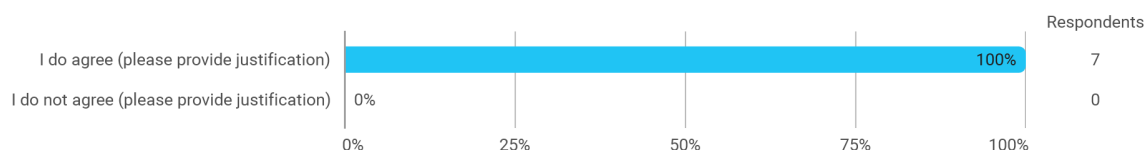


Disagreement justifications:

- This appears to already be clearly communicated in the test guideline in paragraph 16.

12. para 31 - The MTC should be better defined and aligned with other aquatic endocrine TGs.

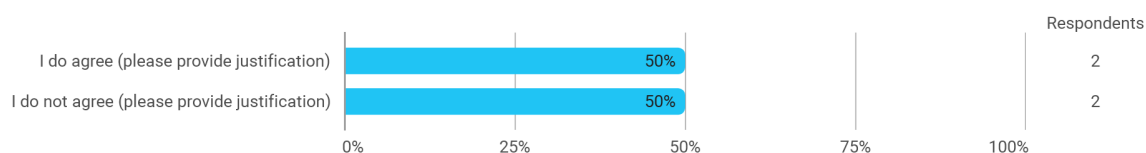
Rational: To avoid regulatory uncertainty and unnecessary repeat or higher tier studies



Agreement justifications:

- Better defined, but MTC must consider TG exposure duration
- see previous comments
- Some additional guidance around the MTC would be good so that we are not expected to kill fish at the top level, but we need to maintain the criteria that the test is still valid if one of the test levels is lost due to systemic toxicity. It is not always possible to avoid any signs of systemic toxicity in the study design and there is pressure to test at high concentrations to make sure that ED is not missed by testing too low, so keeping the ability to drop the top dose and still have a valid study is helpful.
- update needed
- see comments on other relevant TGs

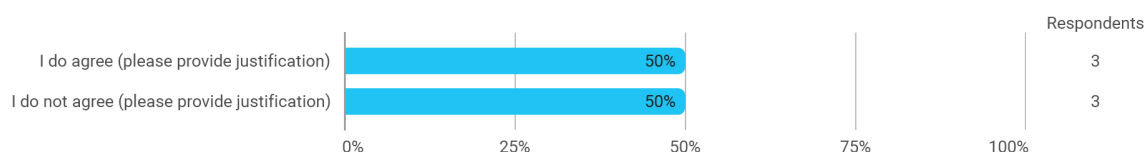
13. para 53 - Genetic sex determination in individual Japanese medaka is based on the presence or absence of the medaka male-sex determining gene, DMY, which is located on the Y chromosome. The genotypic sex of medaka can be identified by sequencing the DMY gene from DNA extracted from for instance a piece of anal fin or dorsal fin. The presence of DMY indicates a XY (male) individual regardless of phenotype, while the absence of DMY indicates a XX (female) individual regardless of phenotype (23). Guidance for tissue preparation and PCR method is given in Annex 9. The genetic sex determination in individual three-spined stickleback is also performed via a PCR method, described in Annex 10. The sex ratio determined during the test has to be based on the phenotypic sex, the genetic sex can be used to determine an impact of the test substance at the individual level.



Disagreement justifications:

- ▶ This appears to already be clearly communicated in paragraph 2. " The presence of a genetic sex marker is a considerable advantage as it increases the power of the sex ratio statistics and enables the detection of individual phenotypic sex reversal."

14. The methodology as described in the current guideline is not suitable to test for effects of virtually insoluble particulate materials (e.g. nano-forms of substances, polymers). Adaption of the guidelines methodology with latest scientific findings is indicated so that valid tests with particulate materials can be performed.



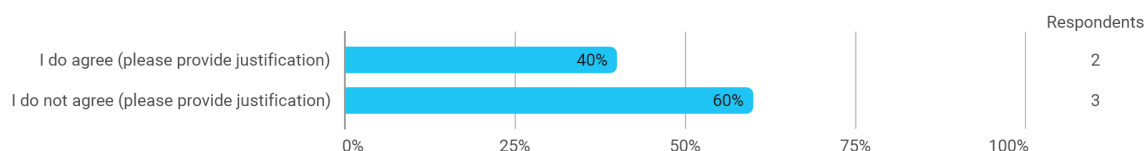
Agreement justifications:

- ▶ guideline should be up to date with current scientific findings
- ▶ reference to guidance document 23 and nano guidance document should be made

Disagreement justifications:

- ▶ See OECD GD 23 and OECD GD for aquatic toxicity testing of manufactured nanomaterials
- ▶ Some general wording to allow for testing a diversity of materials could be included (i.e. for testing nanomaterials or polymers above water solubility limit etc.); however detailed instructions on testing difficult substances should not be addressed in individual toxicity testing guidelines, but should be preferentially addressed in guidance documents such as OECD 23.
- ▶ Methodology is not available. This should be discussed in a separate guideline.

15. It is proposed to update the test criteria to state 'Both male and female fish should meet the criteria of being within the 30 – 70% range'



Agreement justifications:

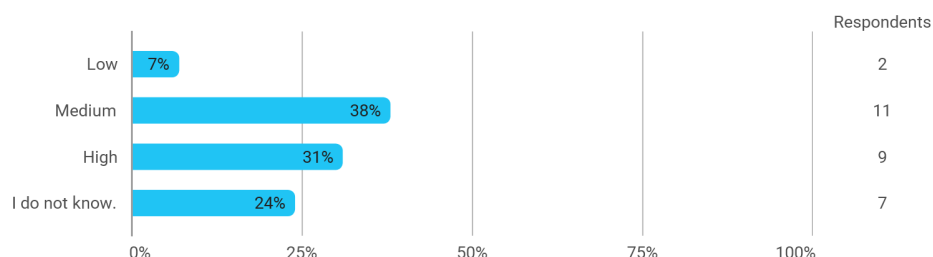
- Too high variance in the control hinder the potential effects of the substance for the sex ratio

Disagreement justifications:

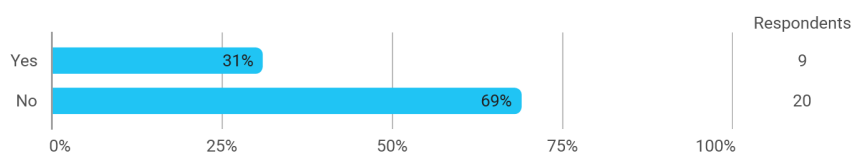
- The guideline is already clear on this (para 16 and annex 2).
- No added value

C.8 Test No. 236: Fish Embryo Acute Toxicity (FET) Test

Relevance of revision

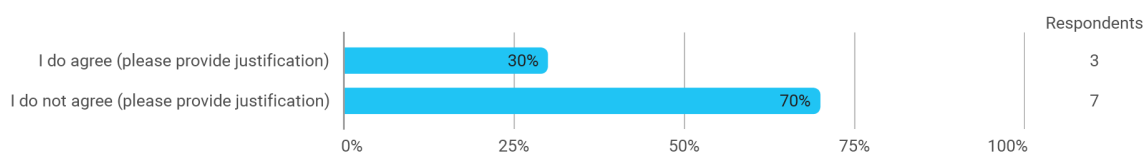


Interest in being part of an OECD Expert Group



The following suggestions have been collected:

- The test could be extended to investigate thyroid disruptors by adding a few specific endpoints like thyroid hormone measurements, thyroid follicle morphology (using transgenic lines or immunostaining) and eye histology**
Rational: thyroid hormones (THs) are essential for early development of all vertebrates, including fish. zebrafish embryo exposure experiments according to TG 236 are ideally suited to assess impact of thyroid disruptors on early development. simple analyses like TH level measurements (ELISA or LC/MS) and thyroid follicle morphology analyses (transgenic lines or immunostaining) can be implemented to assess direct impact on the target system. eye development is regulated by THs and can be assessed as apical, population-relevant endpoint by simple histological analyses of the retinal layers. See ongoing work in EU project ERGO and the Tender project "iFEDT"



Agreement justifications:

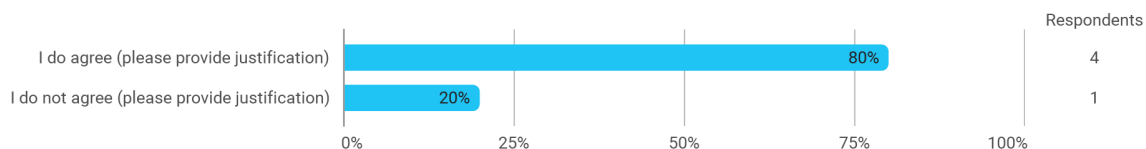
- ▶ strongly agree
- ▶ Update needed when enough background knowledge would have been gathered on these endpoints

Disagreement justifications

- ▶ Xeta already does the job.
- ▶ A number of questions need to be discussed before decisions on the inclusion of thyroid-specific endpoints can be made: Would this increase the number of animals in the test? How conclusive can these endpoints be? Would a definite result on thyroid-related activity be possible or would further testing be required? Moreover, using transgenic lines of fish is not just an addition of endpoints but a major change to the test. How would this affect the overall number of fish used for TG 236 testing? If heterozygous fish are used for breeding, there will be a considerable proportion of non-transgene offspring which could not be used for this purpose and would probably be killed. Therefore, homozygous lines would be required for animal welfare reasons.
- ▶ there are other more suitable TGs in draft for these endpoints. OECD 236 was designed to reduce the number of OECD 203 studies.
- ▶ Disagree. The test guideline has been validated and standardized as a surrogate for the acute fish test with a clear focus on deriving a LC50. Before other endpoints can be added this would need appropriate data bases and ringtesting. Currently other test guidelines are under development using genetically modified zebrafish embryos. Perhaps there is a better chance to still incorporate additional endpoints. For the FET the focus should still be to have an alternative method for OECD TG 302. This is only the case if the endpoint is focused on the LC50.
- ▶ Methodology not yet established, besides, the goal of this study is to assess acute toxicity and proposed endpoints relate to sub-lethal effects. It is difficult to measure both with the same range of concentrations.
- ▶ These endpoints are not part of a standard FET assay, and may warrant a separate TG.
- ▶ Such modifications would be best handled as a separate [new] TG

2. Include a section on initial considerations in line with the latest update of OECD TG 203 (18 June 2019).

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

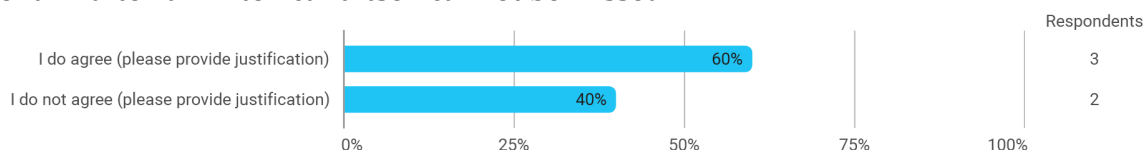
- Align with TG 203
- Need to align with current TG 203

Disagreement justifications

- Disagree. Could be done when the guideline is anyhow revisited. No urgent need.

3. **para 6 - The sentence 'For instance, the protoxicant allyl alcohol (9) has been missed in the FET.' should probably read: 'For instance, the toxicity of the protoxicant allyl alcohol (9) ... has been missed.'**

Rational: Editorial - A toxicant itself cannot be missed.



Agreement justifications:

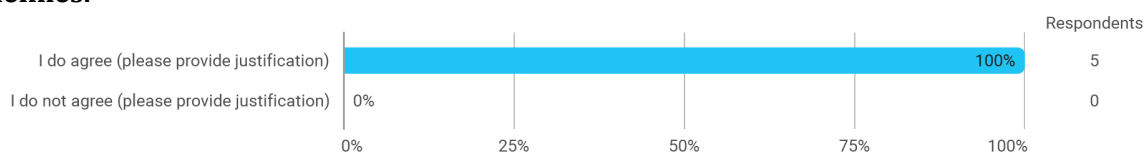
- reads better

Disagreement justifications

- Disagree. Could be done when the guideline is anyhow revisited. No urgent need.
- Small edit, not worth changing

4. **para. 10: add the following text from OECD TG 203: When testing difficult test chemicals, Guidance Document No. 23 (OECD, 2019) should be consulted, and the study design modified appropriately. For volatile and other difficult test chemicals, further specific measures should be taken (OECD, 2018). Any silicone materials that were in contact with the test solution(s) should be preferably discarded and not reused for subsequent tests with different test chemicals.**

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

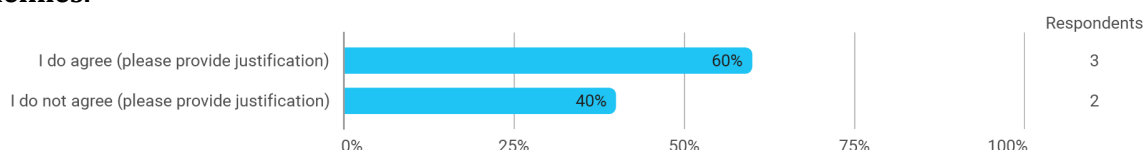
- Align with TG 203

- The problems with especially volatile and adsorbing compounds unfortunately have not been tackled as needed so far. The reference to OECD GD 23 will not totally solve this problem. If the guideline is adopted the topic should be adequately handled. In case of volatile or adsorbing compounds the methodology has known weaknesses. This needs to be taken up.

- Important to link to TG 23 guidance

5. para. 12: align the wording with OECD TG 203. It is proposed to use the following wording from OECD TG 203: If natural water is used, DOC or TOC and nitrate-content (NO₃) should be measured once prior to the test.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

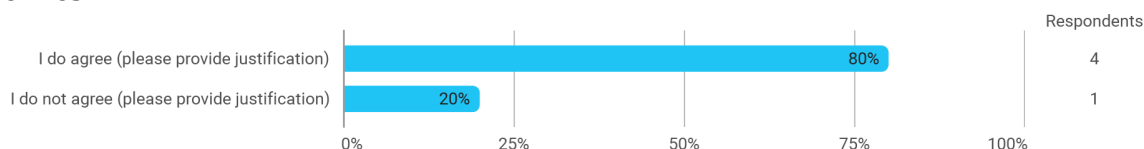
- Align with TG 203

Disagreement justifications

- CROs are doing periodic checks of the DOC/TOC and nitrate but not before each test. Thus, this could be added but is not urgently needed.
- Need to carefully consider "and" vs "or" in this sentence. Robust discussion needed

6. para. 13: to be replaced by Para. 17 from OECD TG 203. The following text from Para. 23 of OECD TG 203 should also be included: Low toxicity solvents only (i.e. acetone, ethanol, methanol, tertiary-butyl alcohol, acetonitrile, dimethyl formamide, dimethyl sulfoxide, and triethylene glycol) as recommended in Guidance Document No. 23 (OECD, 2019) should be used whilst solvents of unknown toxicity should not be used. It should be noted that in spite of the low toxicity for fish, dimethyl formamide and dimethyl sulfoxide should be avoided where possible on human health and safety grounds.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

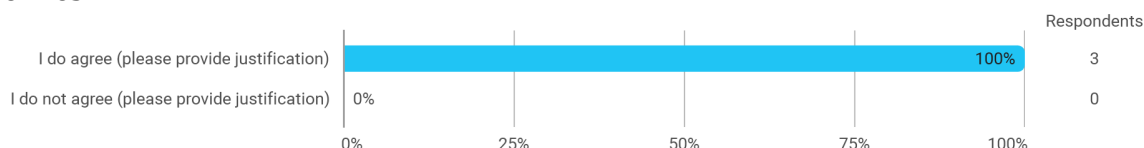
- Align with TG 203
- Agree, in alignment with TG 203

Disagreement justifications

- DMF is a routinely used solvent in aquatic testing. It has been shown to be more suitable in most cases as e.g. acetone or ethanol. In general, untested substances, as routinely tested in ecotox labs, have to be handled as if they were CMR compounds. Thus at least the safety measures in the labs should always be appropriate. Based on the fact that DMF was predominantly used by CRO's most of the historical data being available for possible effects of solvents, relate to DMF. The currently ongoing project investigating whether always a solvent control and a control is needed when a solvent is used in a study based their evaluations as well on a high number of studies which have been performed with DMF. The hope is, that this evaluation will allow to skip the negative control and by that reduce animal numbers in vertebrate tests. If we lose DMF that might have a negative impact on this decision.

7. Add Para. 18 from OECD TG 203

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



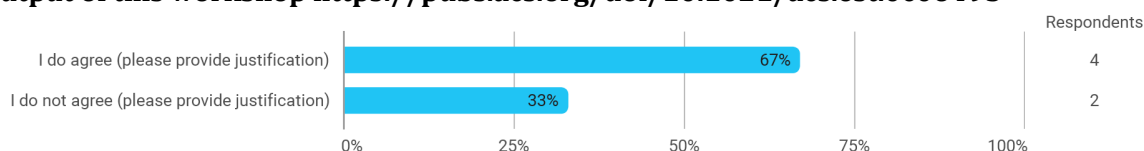
Agreement justifications:

- Align with TG 203

8. paras 28, 34 - Since the use of the zebrafish embryo toxicity is already a basis to study additional endpoints, which are not related to acute toxicity, the extension of the observation, specifically observation on behaviour could be already included. One simple endpoint is for instance the touch evoked response. The test is easy to perform, by touching the embryos at stage 96 hpf with pipette tip or fine brush in the trunk region. Controls are showing a fast response, namely swimming away from the tactile stimulus.

Rational: The additional information can be used to identify chemicals that are having an influence on the nervous system - there recent activities on the level of OECD for developmental neurotoxicity and an zebrafish DNT interest group is working on an assay for the assessment of DNT by using behavioral endpoints (locomotor response assay - light dark transition test). The use of the zebrafish embryo as an alternative test system within an in-vitro based guidance document for DNT testing is planned to be submitted to the OECD WNT end of 2021. Further I did participated a workshop at the UBA in 2019 - "The behaviour of non-target organism after exposure to chemicals - Possibilities of implementation in the regulatory process", where we discussed how to implement the endpoint behaviour of non-target organism in existing OECD

guidelines, rather than setting up new guidelines the idea was discussed to include the additional behaviour endpoint in existing guidelines. A recent publication in ES&T was an output of this workshop <https://pubs.acs.org/doi/10.1021/acs.est.0c06493>



Agreement justifications:

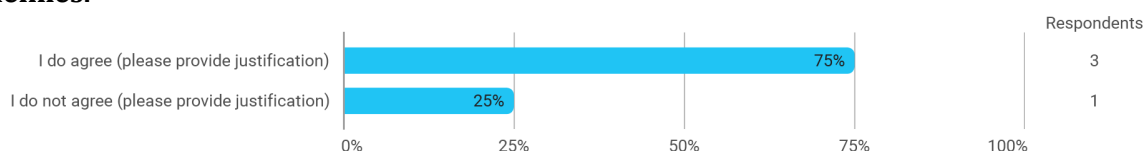
- The inclusion of the touch evoke response would help identifying neurotoxicants in the TG 236. However, discussion with Stefan Scholz (UfZ Leipzig) is needed to decide on its inclusion based on the most recent findings.
- this would be a useful addition at not much extra effort.

Disagreement justifications

- Disagree. The test guideline has been validated and standardizes as a surrogate for the acute fish test with a clear focus on deriving a LC50. Before other endpoints can be added this would need appropriate data bases and ringtesting. Currently other test guidelines are under development using genetically modified zebrafish embryos. Perhaps there is a better chance to still incorporate additional endpoints. For the FET the focus should still be to have an alternative method for OECD TG 302. This is only the case if the endpoint is focused on the LC50.
- While possible, this is not a needed inclusion for the FET assay and was not included in the validation study. I am opposed to including additional endpoints (such as TER or the thyroid endpoints) without robust validation

9. Analytical measurements (Para. 35-36): align with Para 24-25 of OECD TG 203

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

- Align with TG 203

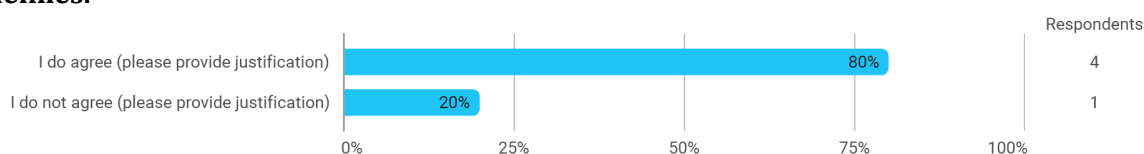
Disagreement justifications

- Disagree Measurements within a FET study need to consider as well practical aspects as the size of test vessels. The Multiwell-plates do not allow the approaches as they are used for e.g. the fish acute test just because of the by far smaller dimensions. Better to keep para 35 as it is. Plates are often covered with a seal. If those needs to be opened or exchanged for measurements a high risk for cross contamination exists. This problem, in addition to the problematic space situation considering the size of standard e.g. oxygen electrodes these measurements are not practical and for a 48 hour study not needed. Might be possible for

semi-static tests with media renewal e.g. after 24 hours. fresh and aged media in a semi-static design .

10. Treatment of results: add the following text from Para. 31 of OECD TG 203: It is recommended that results should be calculated using the measured concentrations of the test chemical. If the deviation from the nominal concentrations is smaller than 20%, results may also be based on the nominal concentrations. It should be noted that it is often useful to have both measured and nominal effect concentrations quoted, see Guidance Document No. 23 (OECD, 2019).

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

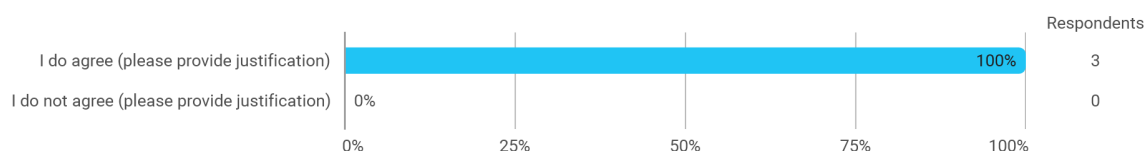
- Align with TG 203
- Important language from TG 203/ TG23

Disagreement justifications

- The information given is sufficient. Different regulatory frameworks use different approaches for the calculations. Harmonization would be appreciated but might not be the right way to do it via the OECD technical guidelines.

11. 11. Include the Annex 3 from OECD TG 203 or at least include consistent information related to the content in particulate matter, TOC and COD in the description what makes an acceptable dilution/test water.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.

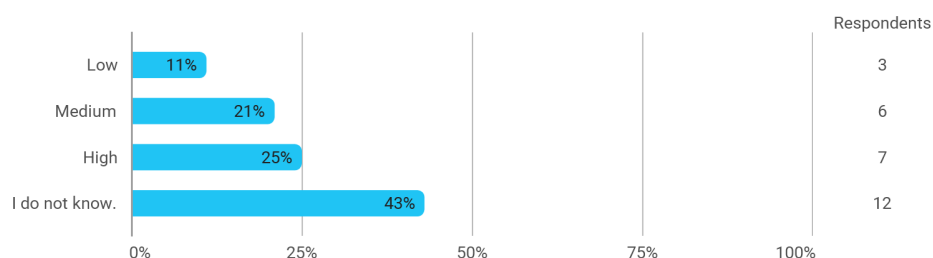


Agreement justifications:

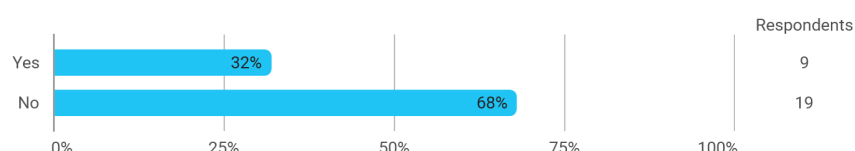
- Align with TG 203

C.9 Test No. 240: Medaka Extended One Generation Reproduction Test (MEOGRT)

Relevance of revision

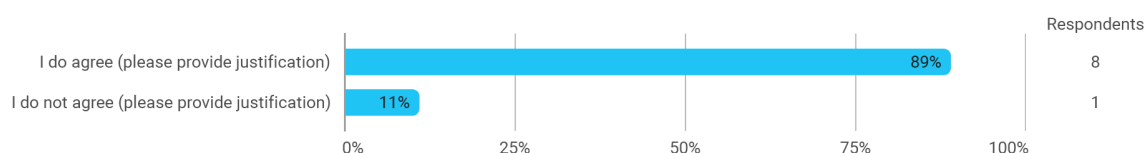


Interest in being part of an OECD Expert Group



The following suggestions have been collected:

- para 3 - This paragraph suggests the TG should be revised in light of experience. It has now been 6 years since publication - are there enough studies to complete this review? Rational: This is a highly complex, animal intensive study with stringent validity criteria. To ensure the TG is practical in light of real use it may be timely to consider if enough data are now available.**



Agreement justifications:

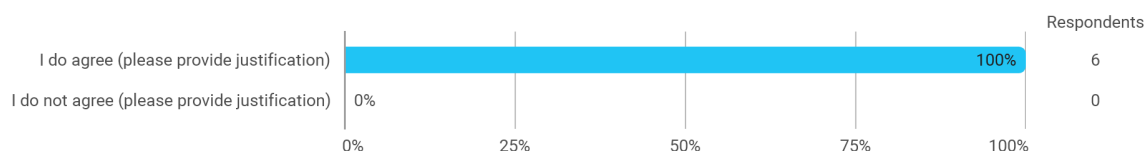
- ▶ Japan is proposing a TG update.
- ▶ It was discussed at the VMG-Eco meeting in October 2021 that a validation report will be prepared by Japan. Based on this, it can be judged whether the TG 240 can be revised.
- ▶ This guideline is not fit for purpose. It hasn't been fully validated and is very difficult to successfully perform.
- ▶ This review should be conducted considering there was never an OECD validation report for this TG, and these studies are now being conducted more routinely. We are planning a data analysis project to assess the performance of the MEOGRT through assessment of historical control data (data from post-validation studies already collected).
- ▶ Strongly agree - this was included in the TG and it would now seem appropriate especially as there is already a project to make some modifications.

Disagreement justifications

- ▶ It is unclear if there is sufficient information (i.e. experience) to revise this guideline. Are there enough studies to complete a review?

2. para 22 - The MTC wording should be aligned with other aquatic endocrine TGs.

Rational: To avoid unnecessary repeat testing.

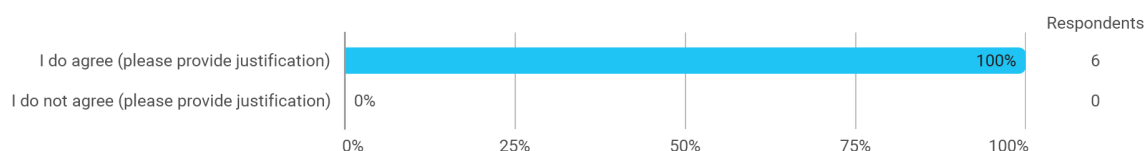


Agreement justifications:

- Update and harmonization needed
- Strongly agree see comments on other relevant TGs

3. para 33- Add wording from TG 210, paragraph 28: “Where deformities and associated abnormal behaviour are considered so severe that there is considerable suffering to the organism, and it has reached a point beyond which it will not recover, it may be removed from the test. Such animals should be euthanised and treated as mortalities for subsequent data analysis.”

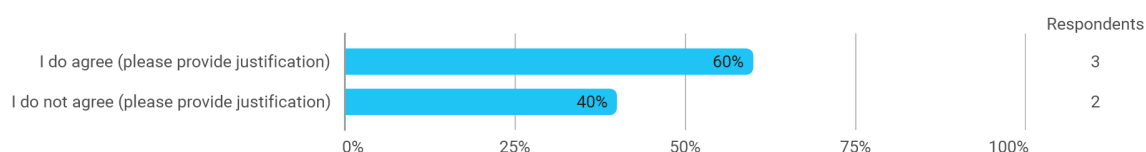
Rational: For reasons of consistency and animal welfare, instructions on removing animals from tests due to severe suffering should be harmonised.



Agreement justifications:

- This change is warranted for animal welfare reasons.
- In line with the 3Rs principles
- agree
- although toxic effects are not expected in ED tests if dose range is correctly designed
- update and harmonization needed

4. The methodology as described in the current guideline is not suitable to test for effects of virtually insoluble particulate materials (e.g. nano-forms of substances, polymers). Adaption of the guidelines methodology with latest scientific findings is indicated so that valid tests with particulate materials can be performed. Furthermore, shortcomings in the validation process were mentioned at e.g. SETAC conferences. Hence, a critical review of the guideline might be warranted.



Agreement justifications:

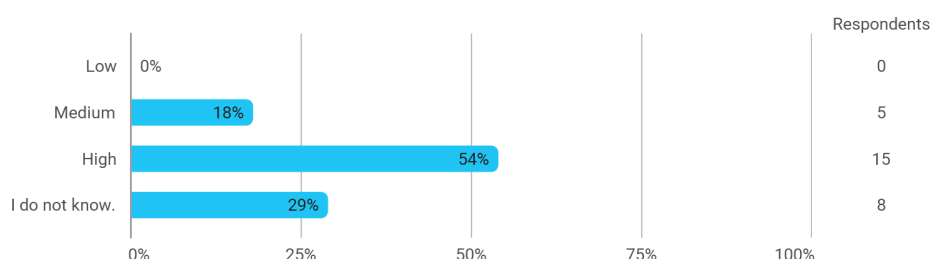
- Validation was inadequate and a critical update including a transparent OECD validation report is required.
- see above

Disagreement justifications

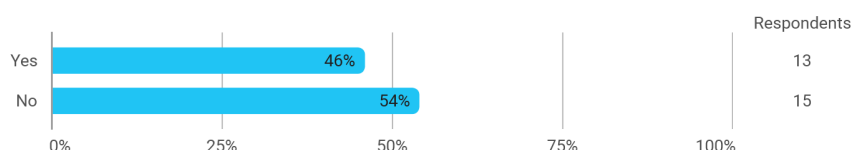
- Some general wording to allow for testing a diversity of materials could be included (i.e. for testing nanomaterials or polymers above water solubility limit etc.); however detailed instructions on testing difficult substances should not be addressed in individual toxicity testing guidelines, but should be preferentially addressed in guidance documents such as OECD 23.
- Methodology is not fully established yet and needs separate guidance.

C.10 Test No. 201: Freshwater Alga and Cyanobacteria, Growth Inhibition Test

Relevance of revision

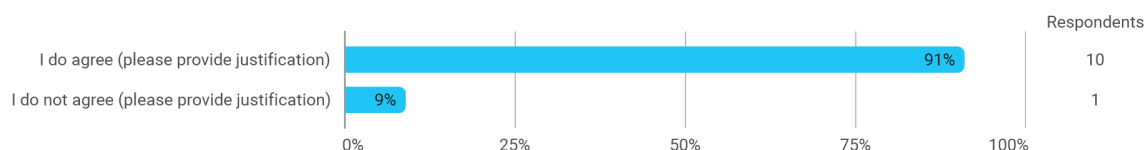


Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. It would be good if marine species of algae were included in this test guidance document. *Phaeodactylum tricornutum* and *Skeletonema costatum* are included in ISO guideline 10253:2016 and OCSPP 850.4500.
Rational: It would be good to have a common OECD guideline to follow for plant protection product regulatory purposes.



Agreement justifications:

- agree. For the CLP purposes of then data from marine species need to be evaluated

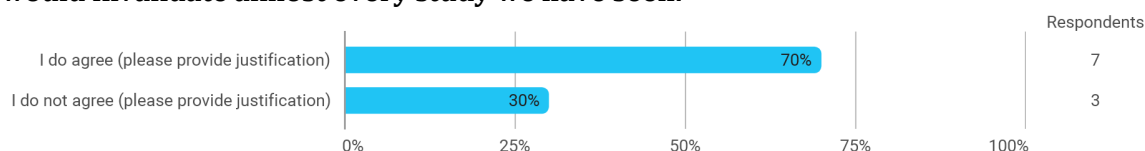
- ▶ Marine ecosystems should be considered in risk assessment and accordingly guidance for a marine algae species should be provided for such an assessment.
- ▶ very good addition, we are currently conducting a review to the EA, UK. The validity criteria for marine species should be considered and altered as appropriate.
- ▶ Marine ecosystems should be considered in risk assessment and accordingly guidance for a marine algae species should be provided for such an assessment.
- ▶ But also the acceptability criteria for those species should be assessed as well.
- ▶ Marine ecosystems should be considered in risk assessment and accordingly guidance for a marine algae species should be provided for such an assessment.
- ▶ *S. costatum* is especially relevant given its wide use in standard toxicity testing

Disagreement justifications

- ▶ Disagree. This guideline is a guideline for freshwater species. Marine species can only be added after a proper standardization and validation process. The Guideline in the current form is not suitable for marine species. Further species should be included only if some kind of ring test can demonstrate that the validity criteria specified in the guideline can be met

2. **paras 1 (and 11)- For risk assessment refinement purposes (e.g., geomean) often studies with non-OECD recommended algae species are submitted. For these species it is often seen that the validity criteria (control treatments) as set out in OECD 201 cannot all be fulfilled. Often the growth factor criterion >16 is fulfilled by large, but then a flattening of the growth curve is observed, and the section-by-section growth rate CV is often observed to be > 35%, while the overall growth rate CV is observed to be < 7% (or 10%). Further guidance on the interpretation of the validity criteria, particularly for the section-by section growth rate CV when species are tested that were not validated (or ring tested) for the OECD guideline, i.e. non-standard test species, is required.**

Rational: Applying the current validity criteria to non-OECD recommended test species would invalidate almost every study we have seen.



Agreement justifications:

- ▶ Important as otherwise the test becomes invalid
- ▶ agree
- ▶ To add to this, species like *Navicula* (diatom) or *Anabaena* which clump while being cultured, it's often difficult to ensure a homogeneous sample is taken, therefore the guideline should advise on some of the details for appropriate culturing and sampling/counting of the different species. The section-by-section specific growth rate requirement for *Anabaena* needs revision.
- ▶ Some clarification is needed, especially on the option suggested in the guideline that shorter exposure time may be used for endpoint calculation (i.e. 0-48h) if all validity criteria are met.

The repetition of studies with non- standard species due to the failure on the criterion for section-by-section growth rate CV, shows that the endpoints are not significantly different. A re-evaluation of the need of this criterion would be welcome

- Species specific validity criteria, especially for slow-growing algae, is warranted.

Disagreement justifications

- Guidance is provided in this regard in OECD 201 (e.g. prolongation of test duration, adjustment of temperature). It should be clear that maintenance of exponential growth is a basic principle of the test. If this aspect cannot be met despite modifications, the species probably is not suitable for this type of test.
- Guidance is included in this regard in OECD 201 (e.g. prolongation of test duration, adjustment of temperature). It should be clear that maintenance of exponential growth is a basic principle of the test. If this aspect cannot be met despite modifications, the species probably is not suitable for this type of test.
- Guidance is included in this regard in OECD 201 (e.g. prolongation of test duration, adjustment of temperature). It should be clear that maintenance of exponential growth is a basic principle of the test. If this aspect cannot be met despite modifications, the species probably is not suitable for this type of test.

3. Currently the following scheme is applied:

A1 The biomass in the control cultures has to increase exponentially by a factor of at least 16. If not, the study is not reliable. In cases where the biomass in the control cultures has increased by a factor of at least 16:

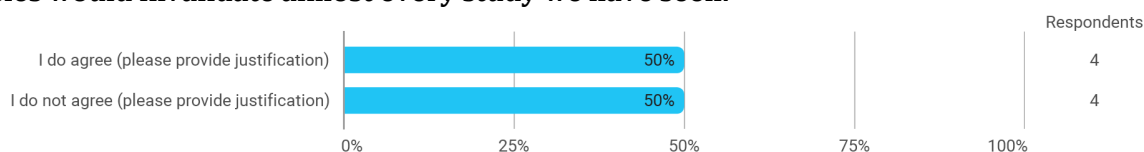
B1 Look at the mean CV for section by section specific growth rate and CV of average specific growth rate. Is the value of the parameter which passes the criterion near the cut-off value or much lower? In case it is much lower, this could support the validity of the study endpoints.

B2 Are the confidence intervals of the ErC50 an EyC50 wide or narrow.

B3 Consider the growth factors of each time frame (e.g., 0-24, 24-48 and 48-72 hours for OECD 201) in combination with that of the whole period. Is growth steady over the time course, or is it decreasing? A flattening growth is often observed, also in the case of standard test species. In that case have a look at the growth factor in the last 24 hours. Is it still reasonable (growth factor > 2.51).

Use all of the above in a WoE approach to accept or reject the exceedance of the mean CV for section by section specific growth rate or CV of average specific growth rate.

Rational: Applying the current validity criteria to non-OECD recommended test species would invalidate almost every study we have seen.



Agreement justifications:

- Agree with the idea that a Weight-of-evidence approach taking into consideration of all relevant parameters should be applied to accept or reject the exceedance of the mean CV for

section by section specific growth rate. However, the language needs to be more clear and organized to reflect this

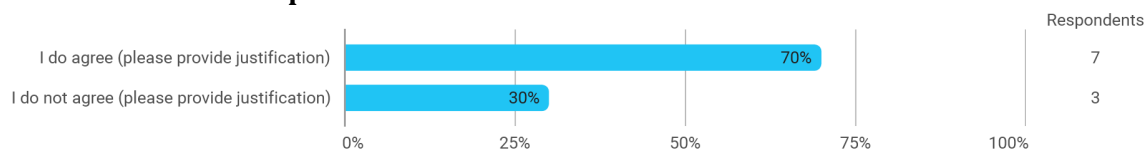
- Species specific validity criteria, especially for slow-growing algae, is warranted.

Disagreement justifications

- Validity criteria (increase cell count, CV section by section growth rate) are combined with the suitability of the data for EC-calculation (width of confidence intervals). These are two completely different, independent aspects.
- Validity criteria (increase cell count, CV section by section growth rate) are combined with the suitability of the data for EC-calculation (width of confidence intervals). These are two completely different, independent aspects.
- The CV criteria have been chosen based on a data evaluation and should only be changed if data demonstrate a need for this. The section by section growth rate is needed to assure exponential growth over time and was added when the focus of the guideline shifted correctly to the growth rate. A comment related to B2: Green algae studies belong to the studies with the best analytical fits among ecotox studies. In most cases the confidence intervals around the EC50 (growth rate and biomass) are very narrow. B3 - this is exactly covered by the section by section growth rate. It would be sufficient to provide the information in the guideline that the section-by-section growth rate CV criterion may not be relevant for species other than the standard green algae. No need to specify any precise approach.
- Validity criteria (increase cell count, CV section by section growth rate) are combined with the suitability of the data for EC-calculation (width of confidence intervals). These are two completely different, independent aspects.

4. para 11 - The second validity criterion, i.e. "The mean coefficient of variation for section-by-section specific growth rates (days 0-1, 1-2 and 2-3, for 72-hour tests) in the control cultures must not exceed 35%" requires adaptation for the cyanobacteria or diatoms.

Rational: It is impossible to reach $\leq 35\%$ while testing anabaena or navicula due to the growth mechanism of those species.



Agreement justifications:

- This is an issue very often. If growth of control and test item treatments are following the same pattern and the other validity criteria are met, this VC seems to be less important for the overall evaluation
- this criteria is the most often failed and one that should be researched especially if adding marine species.
- Also clarification for processing of samples.

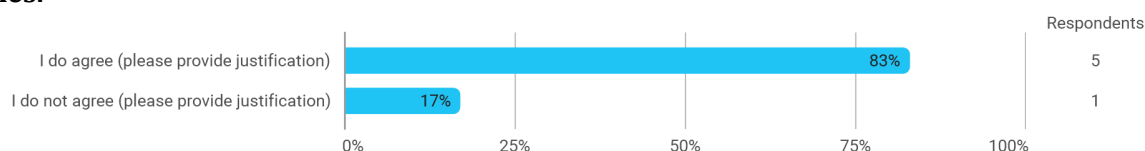
- For diatoms and cyanobacteria that might be correct. For the standard green algae the criteria is correct.
- Species specific validity criteria, especially for slow-growing algae, is warranted.

Disagreement justifications

- Although *Navicula* and *Anabaena* are more difficult organisms to handle, a mean coefficient of variation for section-by-section specific growth rates below 35% is not an inherent issue. For *Synechococcus leopoliensis* on the other hand we made the same experience (very difficult to meet the validity criteria) but that may just be the lack of experience with this species. Besides, this criteria is a measure for steady growth, which is essential for this test, thus modification should not be done by a whim.
- Although *Navicula* and *Anabaena* are more difficult organisms to handle, a mean coefficient of variation for section-by-section specific growth rates below 35% is not an inherent issue. For *Synechococcus leopoliensis* on the other hand we made the same experience (very difficult to meet the validity criteria) but that may just be the lack of experience with this species. Besides, this criterion is a measure for steady growth, which is essential for this test, thus modification should be considered carefully.
- Although *Navicula* and *Anabaena* are more difficult organisms to handle, a mean coefficient of variation for section-by-section specific growth rates below 35% is not an inherent issue. For *Synechococcus leopoliensis* on the other hand we made the same experience (very difficult to meet the validity criteria) but that may just be the lack of experience with this species. Besides, this criterion is a measure for steady growth, which is essential for this test, thus modification should be considered carefully.

5. Include a section on initial considerations in line with the latest update of OECD TG 203 (18 June 2019).

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

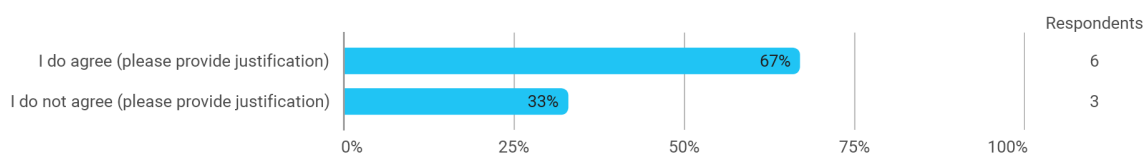
- agree
- Align with other TGs
- also to include some considerations on test items that are likely to be absorbed by algae cells and also centrifugation effects.

Disagreement justifications

- Not relevant to algal testing

6. **para 14 - Add the following text from OECD TG 203: When testing difficult test chemicals, Guidance Document No. 23 (OECD, 2019) should be consulted and the study design modified appropriately. For volatile and other difficult test chemicals, further specific measures should be taken (OECD, 2018). Any silicone materials that were in contact with the test solution(s) should be preferably discarded and not reused for subsequent tests with different test chemicals.**

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

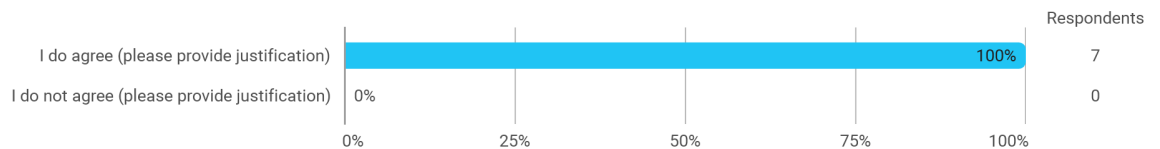
- ▶ agree
- ▶ Align with other TGs
- ▶ also to include considerations for centrifuging vessels used which may affect the analytical results.
- ▶ Reference to TG 23 is useful

Disagreement justifications

- ▶ OECD 23 is referred to in section 13. Silicone materials usually are not necessary to be used in algae tests (contrary to fish tests).
- ▶ Disagree This comment is not fully relevant in case of algae tests as those are designed as static studies and cannot be adapted. In addition, the OECD GD 23 should always be considered when testing difficult substances.
- ▶ OECD 23 is referred to in section 13. Silicone materials usually are not necessary to be used in algae tests (contrary to fish tests).

7. **para 20 - Clarify that if natural water is used it should be justified. Also appropriate information should be provided to demonstrate that this medium is free from endogenous algal populations (including picophytoplankton). It is also needed to align the wording with OECD TG 203: If natural water is used, DOC or TOC and nitrate-content (NO₃) should be measured once prior to the test.**

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.

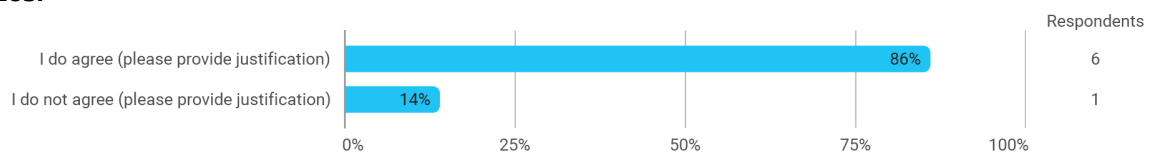


Agreement justifications:

- ▶ agree
- ▶ Align with other TGs
- ▶ it should be included that the cultures should be growing in this modified media before used in the test.
- ▶ No answer for this question
- ▶ Discussion on the use of natural waters warrants robust discussion. I am not fully aligned with all of the suggestions listed here

8. **paras 25 and 28 - To be aligned with Para. 17 from OECD TG 203. The following text from Para. 23 of OECD TG 203 should also be included: Low toxicity solvents only (i.e. acetone, ethanol, methanol, tertiary-butyl alcohol, acetonitrile, dimethyl formamide, dimethyl sulfoxide, and triethylene glycol) as recommended in Guidance Document No. 23 (OECD, 2019) should be used whilst solvents of unknown toxicity should not be used. It should be noted that in spite of the low toxicity for fish, dimethyl formamide and dimethyl sulfoxide should be avoided where possible on human health and safety grounds**

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

- ▶ agree
- ▶ Align with other TGs
- ▶ however some solvents should be avoided (i.e. acetone) as it seems bioavailable and can culminate in excess biofilm growth
- ▶ Reference to TG 23 is useful

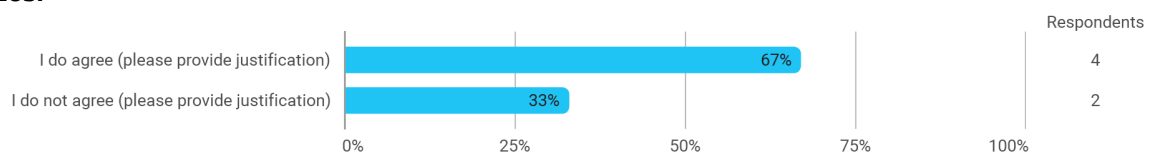
Disagreement justifications

- ▶ DMF is a routinely used solvent in aquatic testing. It has been shown to be more suitable in most cases as e.g. acetone or ethanol. In general, untested substances, as routinely tested in ecotox labs, have to be handled as if they were CMR compounds. Thus at least the safety measures in the labs should always be appropriate. Based on the fact, that DMF was

predominantly used by CRO's most of the historical data being available for possible effects of solvents, relate to DMF. The currently ongoing project investigating whether always a solvent control and a control is needed when a solvent is used in a study based their evaluations as well on a high number of studies which have been performed with DMF. The hope is, that this evaluation will allow to skip the negative control and by that reduce animal numbers in vertebrate tests. If we lose DMF that might have a negative impact on this decision.

9. Add Para. 18 from OECD TG 203.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

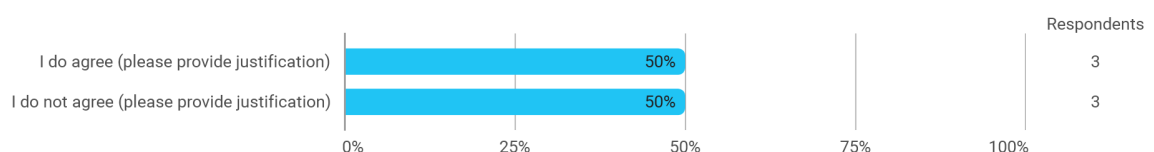
- Align with other TGs
- Agree, but the range for the adapted pH should be more related to the tolerances of the algae species and for some species could be accordingly broader than 6.0 to 8.5

Disagreement justifications

- Different pH range is needed at least. If pH of the stock solution is to be manipulated, then preferably as currently to the pH value of the dilution test water.
- para.18 of OECD 203 indicates that the test should be carried out without adjustment of pH, however modification of the growth media may be necessary for certain purposes, e.g. when testing metals and chelating agents or testing at different pH values.

10. Frequency of analytical determinations and measurements (Para. 33-40): align with Para 24-25 of OECD TG 203.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

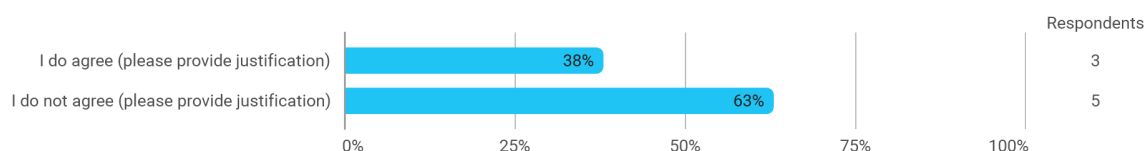
- Align with other TGs

Disagreement justifications

- ▶ para 24 okay. para 25 not acceptable for algae tests: OECD 203 has 1 replicate per treatment, OECD 201 at least 3 replicates. Impact on workload for a single test would be massive and data generated is not very meaningful. For degrading substances with additional samplings maybe the pH can be measured in addition by treatment (not per replicate).
- ▶ 203 describes semi-static renewal exposure which is not possible with algae. Also para 38 needs to clarify the “identical” treatment of each analysis. i.e. Day 0 samples should not be inoculated with algae cells and therefore these samples should not require centrifuging.
- ▶ Disagree to align with paragraph 25. O2 measurement are not needed, pH measurement does not need to be daily in all test vessels. Agree to align with para 24 with respect to the chemical analysis.

11. para 40 - Add the following text from Para. 31 of OECD TG 203: It is recommended that results should be calculated using the measured concentrations of the test chemical. If the deviation from the nominal concentrations is smaller than 20%, results may also be based on the nominal concentrations. It should be noted that it is often useful to have both measured and nominal effect concentrations quoted, see Guidance Document No. 23 (OECD, 2019).

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

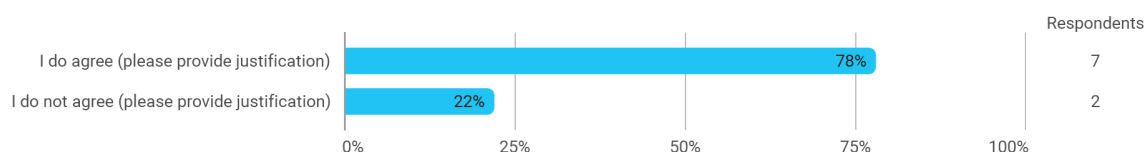
- ▶ Align with other TGs

Disagreement justifications

- ▶ if concentrations are in the range of nominal it should be justified to use only these endpoints. The actual measured concentrations (if 80-120% of nominal) are highly impacted by minor inaccuracies and are not meaningful.
- ▶ Already included in OECD 201, see para 39.
- ▶ Generating both nominal and measured values could cause confusion with too many endpoints being reported. We agree that some clear guidance is needed for the occasions where the test item is degrading quickly.
- ▶ Disagree. The information given is sufficient. Different regulatory frameworks use different approaches for the calculations. Harmonization would be appreciated but might not be the right way to do it via the OECD technical guidelines.
- ▶ Already included in OECD 201, see para 39.

12. para 29 - "Temperature in the range of 21 to 24 ° C, controlled at $\pm 2^{\circ}\text{C}$ " Suggested update: If possible, clarify if the temperature must be within the specified range or if for example $21\pm 2^{\circ}\text{C}$ could be acceptable.

Rational: Incubation conditions would be better specified



Agreement justifications:

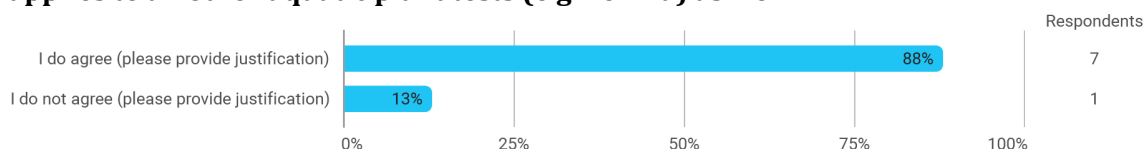
- Agreed, wording is not clear.
- Agreed. The description led to different interpretations in the past.
- Agreed, wording is not clear.
- More precise language is needed

Disagreement justifications

- this seems fine and works well. Labs have long established methods and this would interfere with routine testing and culturing.
- Different species may grow better at different temperatures; suggestion to propose the optimal temperature for each species, also the same should apply for light intensities. All the information can be provided in tabular form.

13. para 30 - "The pH of the control medium should not increase by more than 1.5 units during the test." Delete or alternatively give a wider range because an increase below 1.5 is very difficult to maintain and does in my opinion not impact the outcome of a study. This applies to all other aquatic plant tests (e.g., Lemna) as well.

Rational: Delete or alternatively give a wider range because an increase below 1.5 is very difficult to maintain and does in my opinion not impact the outcome of a study. This applies to all other aquatic plant tests (e.g. Lemna) as well.



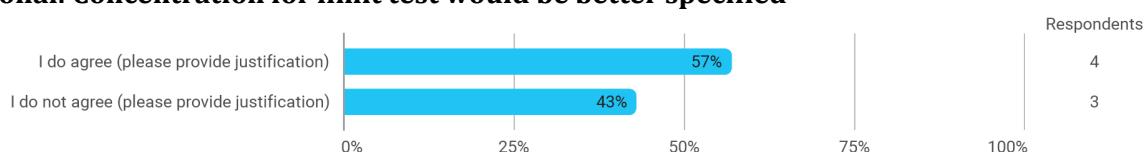
Agreement justifications:

- yes, very important, as authorities are very strict with guidance, but scientific impact is very low, if e.g. growth of algae and test substance recoveries are as expected
- Would suggest including sentence as per OECD 221 (However, deviation of more than 1.5 units would not invalidate the test when it can be shown that validity criteria are met.).
- The pH increase is increasing in parallel to the exponential growth. Thus, the highest increase is occurring towards the end of the study. While the effects normally occur already earlier. Thus, the impact of a pH shift exceeding 1.5 units in most cases is minor and should not be a reason for invalidating studies directly. More information and clarification could thus be helpful.

- Would suggest including sentence as per OECD 221 (However, deviation of more than 1.5 units would not invalidate the test when it can be shown that validity criteria are met.).
- This warrants consideration. pH drift is difficult, if not impossible, to control in algal tests
- Suggestion 13 - Disagreement justifications
- this needs to be clarified, 1.5 unites after 72 hr of growth for green algae is usually achievable, however after 96 hr the exponential algae growth will increase the pH levels up to 2-3 unites and this should not invalidate the study.

14. para 42 - Limit test: In the case of tests with formulated products, it would be appropriate to specify whether the concentration refers to the product or to the active ingredient (or one of the active ingredients).

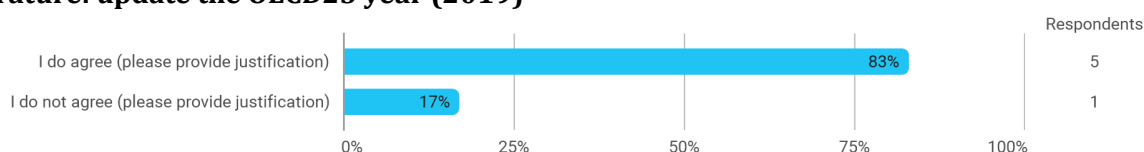
Rational: Concentration for limit test would be better specified



Disagreement justifications

- Generally, it is 100 mg test item/L and accordingly if the test item is a formulation, it is 100 mg formulation/L. Details may depend on the purpose of the study but this document is not the right place to give guidance in this regard (not an algae test system specific question).
- This is not needed since this is dependent on the way the endpoints are used and the regulatory framework
- Generally, it is 100 mg test item/L and accordingly if the test item is a formulation, it is 100 mg formulation/L. Details may depend on the purpose of the study but this document is not the right place to give guidance in this regard (not an algae test system specific question).

15. Literature: update the OECD23 year (2019)



Agreement justifications:

- agree
- Yes, but not urgent.
- Citation needs updating

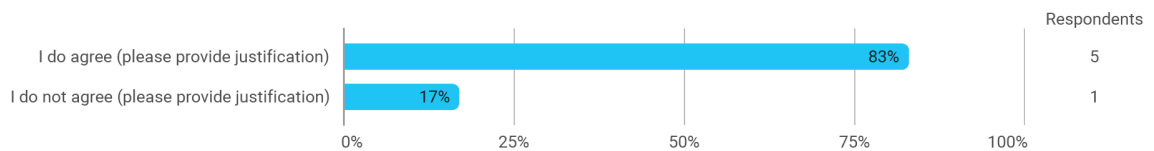
Disagreement justifications

- to be discussed: the OECD 201 states "geometric mean" for evaluation, but this formula is only in the OECD 23 of 2000.

16. Guidance or a cut-off value should be provided for situation when LOEC is set at a concentration with a very low reduction/inhibition.

Rational: Sometimes when the culture performs very well, the variation in the control is so low that even a minute reduction/inhibition (e.g. 0.5%) can be statistically

significant moving NOEC to very low (unrealistic) values while, e.g. EC10 is much higher.



Agreement justifications:

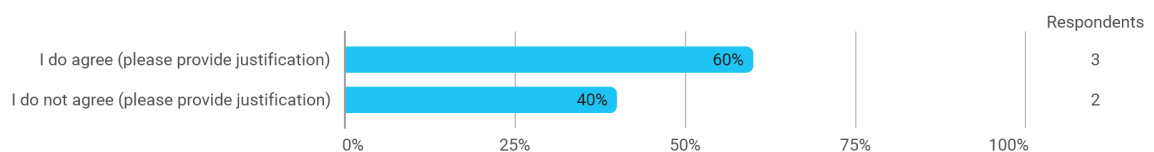
- ▶ sometimes an inhibition of 0.5% can be statistically significant but its biological significance is doubtful.
- ▶ Fully agree. The studies with green algae are frequently differentiating already very minor effects which are biologically meaningless. Therefore, this would be helpful. Currently a good study performance is resulting in very low effect values. Which could result in LOEC values around 1-2 % difference to the controls. This is, considering algae growth under environmental conditions, meaningless.
- ▶ I would suggest re-evaluating guidance of NOECs, LOECs, and EC10s

Disagreement justifications

- ▶ if statistical evaluation determines significance, then it should be reported as such. The whole concept of LOEC/NOEC needs to be discussed, but by now results shouldn't be manipulated in such a manner.

17. Some guidance on the reliability of the ECx endpoints should be added; the clearest guidance to be suggested is what's mentioned in OECD 210 p.32 If an ECx is to be reported, the design of the experiment and selection of regression model should permit estimation of ECx so that (i) the 95% confidence interval reported for ECx does not contain zero and is not overly wide, (ii) the 95% confidence interval for the predicted mean at ECx does not contain the control mean (iii) there is no significant lack-of-fit of regression model to the data.

Rational: At this moment no guidance is provided regarding reliability of the obtained ECx values.



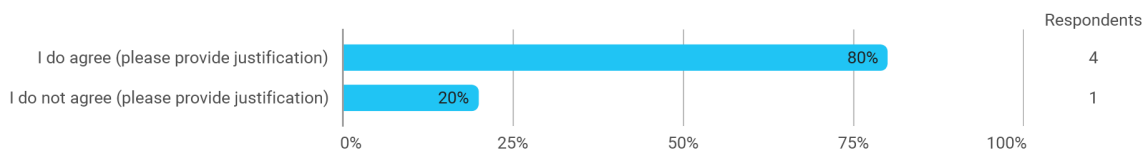
Disagreement justifications

- ▶ Guidance on reliability of ECx values are highly appreciated, but not as in OECD 210. Steepness of curve, normalised width, overlapping of CI are better parameters to determine reliability (rating)
- ▶ Disagree For most algae studies this is not a problem. The CI's around the EC50(s) – which is the main endpoint of this study type, are narrow, especially for green algae.

18. para 47 - The paragraph addresses the controversy of the ErC50 vs the EyC50, which in itself is appreciated. However, it causes confusion with regard to certain publications, where the ErC50 has been compared to mesocosm results and found to

not always be protective. Can the paragraph be revised in the light of these scientific findings?

Rational: This is for a long time a topic of discussion in the risk assessment world and would need to be addressed by both the study guidelines and the risk assessment guidance.



Disagreement justifications

- The comment with respect to comparison with mesocosm studies is wrong! The mentioned publication to which this comment is referring to, is an oral presentation from UBA at the SETAC Rome. This was never published in a peer reviewed journal. Publications in peer reviewed journals as .e.g. by Arts & van Wyngarden came to a different result. The underlying data from the UBA evaluation were never made publicly available. E.g. the aquatic ad hoc group under the umbrella of CLE asked for the data but only received the presentation slides which do not contain the necessary information. The statement that the tier 1 risk assessment is not protective therefore is in contradiction to real publication and other presentations at SETAC where it was presented that the tier 1 risk assessment for primary producers is protective in most cases. Therefore, this comment is wrong and misleading! Anyhow this is a statement which should not be part of a guideline at all. A Guideline describes a test methodology. The regulatory framework is not part of a guideline

19. Navicula pelliculosa, UTEX 664, You can no longer obtain this strain of algae. A new strain should be found and full validated for the method to ensure it can meet the validity criteria.

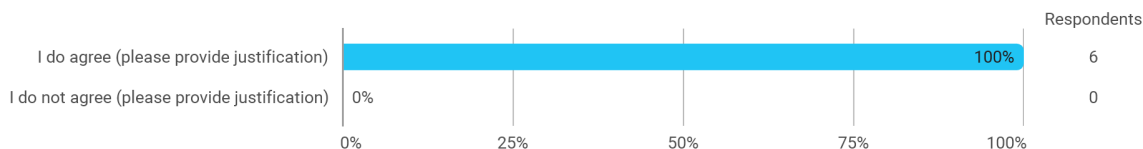
Rational: You can no longer obtain this strain of algae

Adaptation of species suitable for the test.

Rational: Navicula testing is causing a lot of issues. It would be appreciated if the species selection is reconsidered in light of feasibility and regulatory impact.

“Strains shown to be suitable for the test”: Suggested update: See request submitted by Japan for the introduction of other Navicula strains (TG201_SPSF_201112_JAPAN).

Rational: “Strains shown to be suitable for the test”: Strain UTEX664 is unavailable.

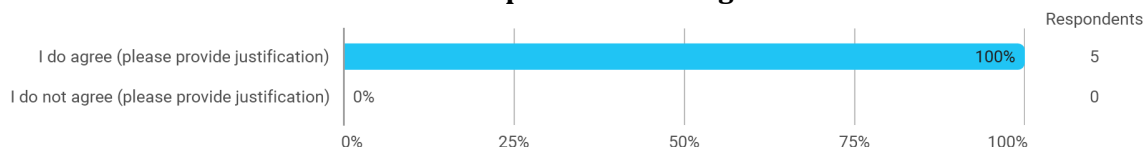


Agreement justifications:

- SPSF already in progress for this change
- agree, this species is difficult to work with and ensure the VC are all met, so this needs to be revised

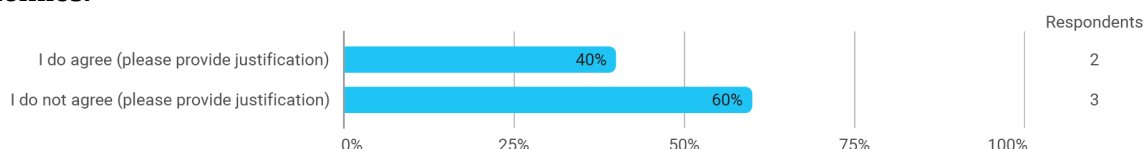
20. For algae species that may form aggregates under growth condition, more detailed guidance on the best practice for handling and cell counting is required.

Rational: At this moment the information provided in the guideline is not sufficient.



21. Include the Annex 3 from OECD TG 203 or at least include consistent information related to the content in particulate matter, TOC and COD in the description what makes an acceptable dilution/test water.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Disagreement justifications

- at least determination intervals need to be provided, e.g. (semi-)annual.
- Composition of algae media is described.
- Disagreed: Media for use in algae testing are described in the guideline.

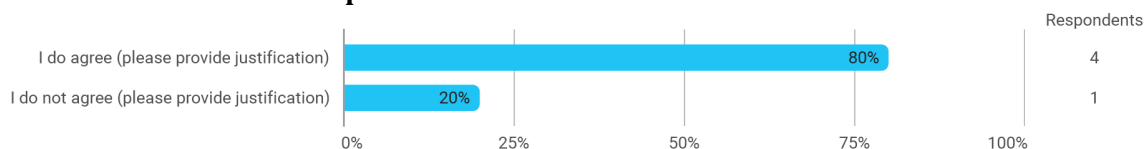
22. Give advice on medium adaptations.

“Growth media”: Pag. 18, Composition of AAP-medium (US. EPA) according to ASTM Suggested update: MgSO₄·7H₂O: 14.7 mg/L instead of 14.6 mg/L.

Rational: “Growth media”: Pag. 18, Composition of AAP-medium Based on the quantities required on page 21 for the preparation (7.350 g / 500 mL), in the medium the content should be 14.7 mg/L and not 14.6 mg/L.

“Growth media”: Page 20, Preparation of OECD medium Suggested update: adding the information “Adjust pH to 8.1± 0.2 with 0.1 N or 1.0 N NaOH or HCl.”

Rational: Annex 3 “Growth media”: Page 20, Preparation of OECD medium: for OECD medium the pH value is indicated only at page 18, without a tolerance range; instead in ISO 8692 the indication for pH is 8.1 ± 0.2.

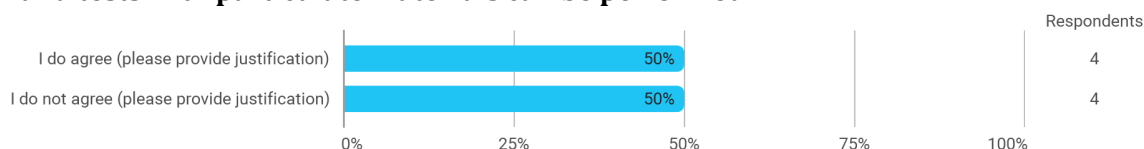


Agreement justifications:

- Agreed: it should be clear that medium adaptation (especially for non-standard species) is possible as long as the validity criteria of the guideline are met
- Align with current best practices

23. The methodology as described in the current guideline is not suitable to test for effects of virtually insoluble particulate materials (e.g. nano-forms of substances, polymers).

Adaption of the guidelines methodology with latest scientific findings is indicated so that valid tests with particulate materials can be performed.



Agreement justifications:

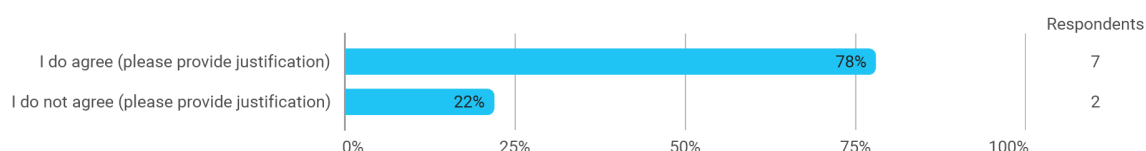
- ▶ very interesting
- ▶ Not algae-specific, rather to be included in OECD 23.

Disagreement justifications

- Not algae-specific, rather to be included in OECD 23.
- ▶ Methodology is not fully established yet and needs separate guidance.
- ▶ The test was standardized and validated for dissolved test items. For difficult substances please refer to the OECD Technical Guidance on difficult substances (OECD GD 23).
- ▶ This is what TG 23 guidance is for

24. Consider if the marine species *Skeletonema costatum* could be included as a test species. If *S. costatum* is not a suitable test species, state this in the test guideline and the reasons why.

Rational: It would be good to have a common OECD guideline to follow for plant protection product regulatory purposes.



Agreement justifications:

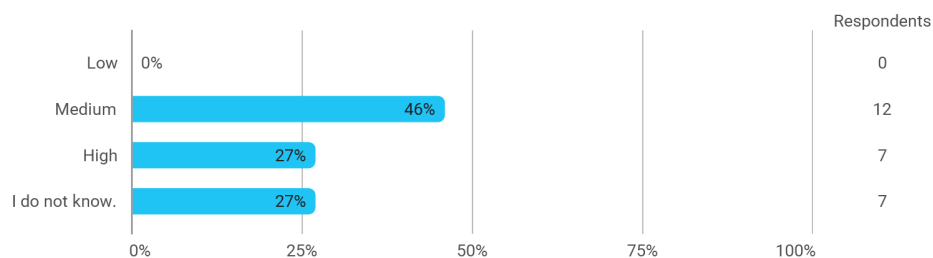
- ▶ see comments for suggestion 1
- ▶ As above, marine ecosystems should be considered in risk assessment and accordingly guidance for a marine algae species should be provided for such an assessment.
- ▶ Marine ecosystems should be considered in risk assessment and accordingly guidance for a marine algae species should be provided for such an assessment.
- ▶ Agreed, this species is routinely used in other jurisdictions

Disagreement justifications

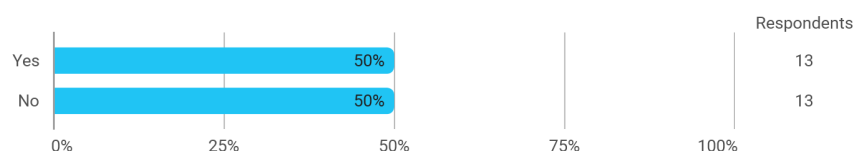
- ▶ Consider a marine species that is suitable and clarify the validity for this species.
- ▶ Disagree. The current guideline is, as stated in the guideline, suitable for freshwater species. The usability for marine species would need standardization and validation.

C.11 Test No. 202: Daphnia sp. Acute Immobilisation Test

Relevance of revision

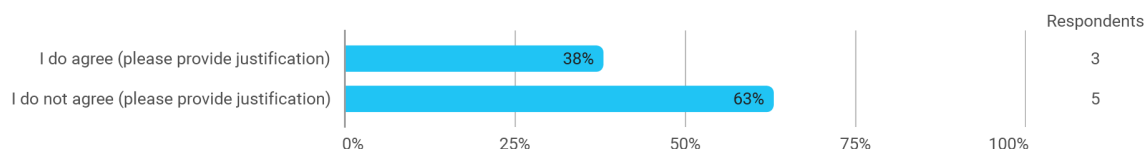


Interest in being part of an OECD Expert Group



The following suggestions have been collected:

- The methodology as described in the current guideline is not suitable to test for effects of virtually insoluble particulate materials (e.g. nano-forms of substances, polymers). Adaption of the guidelines methodology with latest scientific findings is indicated so that valid tests with particulate materials can be performed.**



Agreement justifications:

- It is indeed a problem to conduct daphnia tests with oily formulations or micelles forming substances
- Align with other TGs

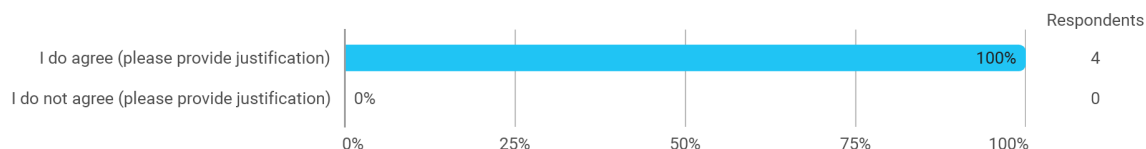
Disagreement justifications

- Not daphnid-specific, rather to be included in OECD 23.
- Methodology is not fully established yet and needs separate guidance.
- The test was standardized and validated for dissolved test items. For difficult substances please refer to the OECD Technical Guidance on difficult substances (OECD GD 23).
- Not daphnid-specific, rather to be included in OECD 23.
- This is what TG 23 is for

- Include a section on initial considerations in line with the latest update of OECD TG 203 (18 June 2019)**

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to

the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.

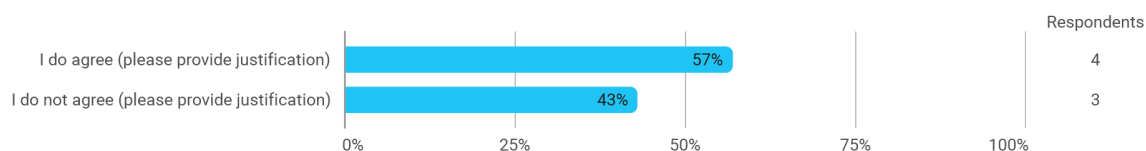


Agreement justifications:

- ▶ nice to have, especially for person not familiar with this test. Further and more importantly chemical producers may be encouraged to deliver such information like water solubility, volatility etc., which are necessary to conduct a good test and avoid repetitions and surprises
- ▶ Align with other TGs
- ▶ but not entirely necessary since there's no need to discuss the animal welfare topic

3. **para 22- It should be recommended to measure dissolved oxygen at least in the lowest concentration in which >50% immobilisation occurred at t=24h in order to assess whether the observed effect is not due to lack of oxygen (caused e.g., by reaction of test item with dissolved oxygen).**

Rational: The purpose of the guideline is to assess the intrinsic toxicity of the test item, without measuring dissolved oxygen it is impossible to assess what caused an effect.



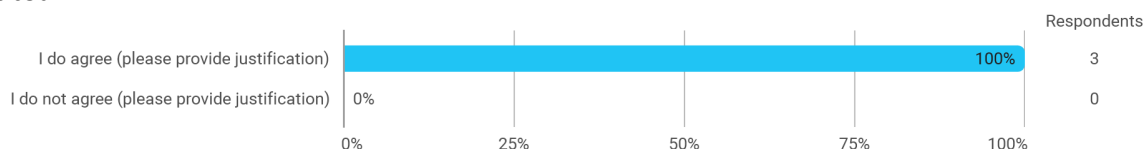
Agreement justifications:

- ▶ That's para 22. Yes, at least. Oxygen availability is a relevant parameter for the mobility of the test organism and accordingly for the interpretation of the test results.
- ▶ Even though this is irrelevant to paragraph 2, also if the daphnia loading is appropriate and the initial DO is well above 60%ASV or 3 mg/L, there is less chance that the DO will drop that much in 24hrs unless the AI reacts with oxygen can cause toxicity. Additional measurements can also be taken by the SD to investigate if this is observed at the RF stage.

Disagreement justifications

- ▶ after 48 h you see if immobilisation was caused by lack of oxygen. Further Oxygen measurement instrument may cause contamination, even after careful cleaning
- ▶ This could be done by the performer already but should not be a must. Would be more suitable for para 22.
- ▶ Not seeing this as a major issue plaguing this assay

4. para 4 and Literature - reference (4) and corresponding text in paragraph 4 should be updated

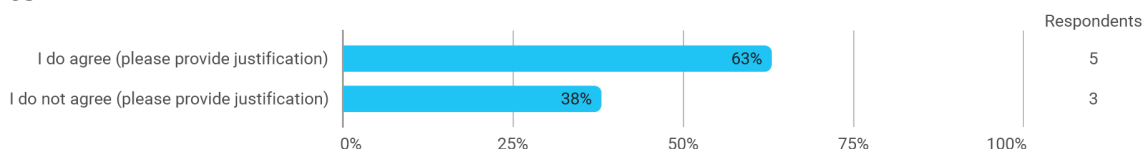


Agreement justifications:

- ▶ e.g. for OECD 23 exist a new version
- ▶ OECD 23 (2019)

5. para 7 - Add the following text from OECD TG 203: When testing difficult test chemicals, Guidance Document No. 23 (OECD, 2019) should be consulted, and the study design modified appropriately. For volatile and other difficult test chemicals, further specific measures should be taken (OECD, 2018). Any silicone materials that were in contact with the test solution(s) should be preferably discarded and not reused for subsequent tests with different test chemicals.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines



Agreement justifications:

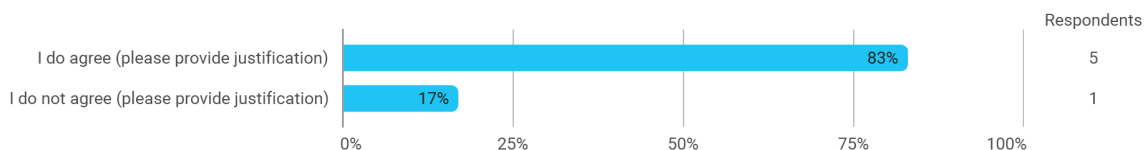
- ▶ more detailed.
- ▶ Align with other TGs
- ▶ Agree
- ▶ Referencing TG23 is useful

Disagreement justifications

- ▶ OECD 23 is referred to in section 4. Silicone materials usually are not necessary to be used in daphnid tests (contrary to fish tests).
- ▶ The reference to OECD TG 23 can be included in the future but is currently no reason for re-opening this guideline.
- ▶ OECD 23 is referred to in section 4. Silicone materials usually are not necessary to be used in daphnid tests (contrary to fish tests).

6. para 10 - Align the wording with OECD TG 203: If natural water is used, DOC or TOC and nitrate-content (NO3) should be measured once prior to the test.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

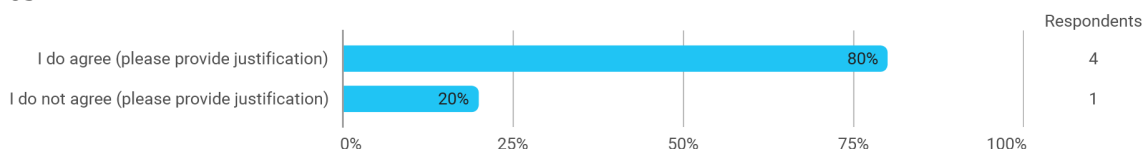
- ▶ NO3 important parameter
- ▶ Align with other TGs
- ▶ however it is also mentioned in para.11
- ▶ Agree on principle, but a robust discussion is needed (especially around the words "and" "or")

Disagreement justifications

- ▶ CROs are doing periodic checks of the DOC/TOC and nitrate but not before each test. Thus, this could be added but is not urgently needed.

7. **para 12 - To be replaced by Para. 17 from OECD TG 203. The following text from Para. 23 of OECD TG 203 should also be included: Low toxicity solvents only (i.e. acetone, ethanol, methanol, tertiary-butyl alcohol, acetonitrile, dimethyl formamide, dimethyl sulfoxide, and triethylene glycol) as recommended in Guidance Document No. 23 (OECD, 2019) should be used whilst solvents of unknown toxicity should not be used. It should be noted that in spite of the low toxicity for fish, dimethyl formamide and dimethyl sulfoxide should be avoided where possible on human health and safety grounds.**

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



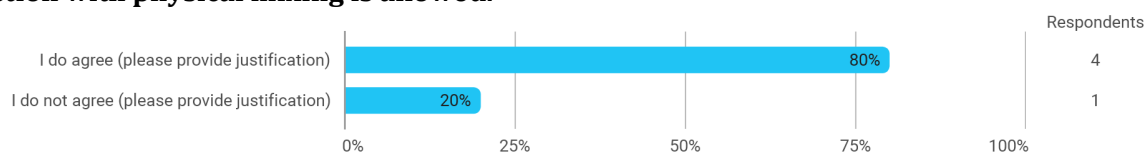
Agreement justifications:

- ▶ It should be recommended, which solvents are truly not toxic to daphnia and the amount of solvent should be mentioned
- ▶ Align with other TGs
- ▶ agree

Disagreement justifications

- DMF is a routinely used solvent in aquatic testing. It has been shown to be more suitable in most cases as e.g. acetone or ethanol. In general, untested substances, as routinely tested in ecotox labs, have to be handled as if they were CMR compounds. Thus at least the safety measures in the labs should always be appropriate. Based on the fact that DMF was predominantly used by CRO's most of the historical data being available for possible effects of solvents, relate to DMF. The currently ongoing project investigating whether always a solvent control and a control is needed when a solvent is used in a study based their evaluations as well on a high number of studies which have been performed with DMF. The hope is, that this evaluation will allow to skip the negative control and by that reduce animal numbers in vertebrate tests. If we lose DMF that might have a negative impact on this decision.

8. **paras 12 and 27 - Chapters should be adapted such, that the use of auxiliary agents such as chemical dispersants or emulsifying auxiliary agents is not not advocated, because of the potential for physical or chemical interactions influencing the apparent toxicity of the test chemical. Appropriate solvents used in conjunction with physical mixing is allowed.**



Agreement justifications:

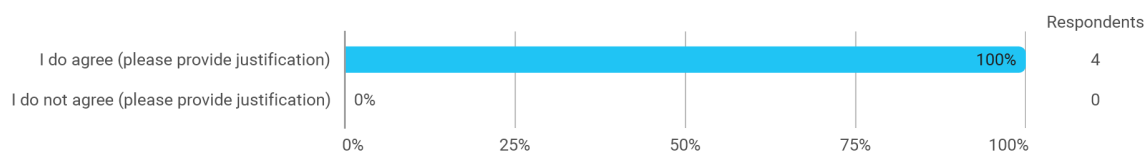
- but then a suggestion should be made for substances which make daphnia stick to the surface
- Align with other TGs
- agree

Disagreement justifications

- Covered in TG23

9. **para 13 - Align with Para. 18 from OECD TG 203.**

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.

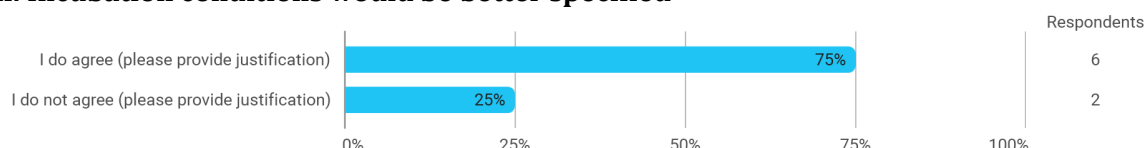


Agreement justifications:

- ▶ absolutely necessary to give clear recommendations especially in context taht the pH should not vary by more than 1.5 units. Further a second test should not be conducted, when the pH is adjusted
- ▶ Align with other TGs
- ▶ Agree

10. para 18 - Sentence “Temperature within the range of 18 and 22 °C, constant within $\pm 1^{\circ}\text{C}$ ” Suggested update: If possible, clarify if the temperature must be within the specified range or if for example $18 \pm 1^{\circ}\text{C}$ could be acceptable.

Rational: Incubation conditions would be better specified



Agreement justifications:

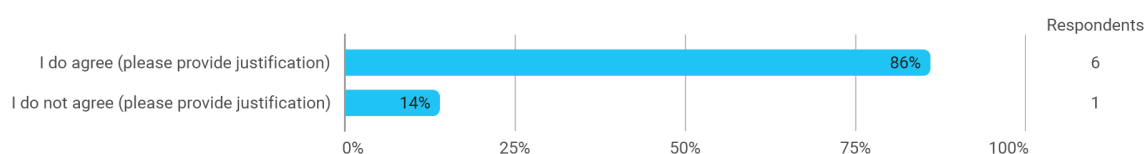
- ▶ Agreed, wording is not clear.
- ▶ agree
- ▶ Agree
- ▶ Agreed, wording is not clear.
- ▶ Agree for clarify

Disagreement justifications

- ▶ it is clear how the temperature should be and allows more flexibility in general temperature than the suggestion.
- ▶ labs have set criteria. to be more specific would disadvantage labs as cultures may be stressed. We test at 20 so altering to say 18 would be a massive jump for our Daphnia culture. as they are very sensitive to these type of changes.

11. para 18 - It would be good to use the same wording as in paragraph 31 of OECD 211 i.e. “The temperature of the test media should be within the range 18-22°C. However, for any test, the temperature should not, if possible, vary by more than 2°C within these limits (e.g., 18-20, 19-21 or 20 - 22°C) as daily range. It may be appropriate to use an additional test vessel for the purposes of temperature monitoring”.

Rational: The description provided in OECD 211 is clearer.



Agreement justifications:

- ▶ Align with other TGs

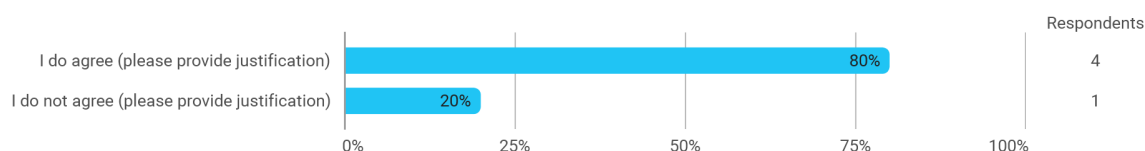
- Agreed, as suggestion #10.
- agree
- Agree
- Agreed, as suggestion #10.

Disagreement justifications

- implies that the temperature can vary by more than 1°C between days. This gives more flexibility for not perfectly temperature regulated labs. But this also requires continues measurement of temperature which was not mandatory regarding the actual version. so the suggestion is okay, but also has disadvantages

12. para 23 - Align with Para. 24-25 of OECD TG 203.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

gives more clear advice, but immediate measurements and stability samples will cause extra costs and problems for some labs due analytical capacities

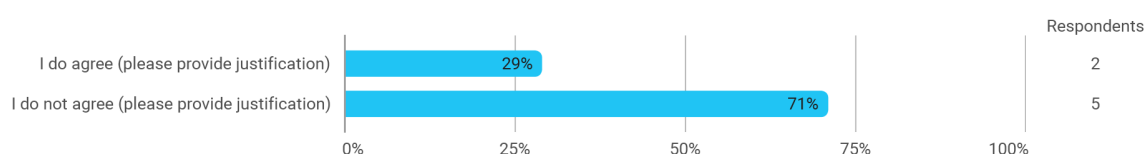
- Align with other TGs
- agree

Disagreement justifications

- disagree to align with paragraph 25 about the frequency of measurement of pH, O2, salinity, etc. It is not necessary to measure them daily in all test vessels. Therefore, the para should stay unchanged. With respect to pH, O2, salinity etc. The currently presented information is sufficient in most cases. Therefore, the adaptation of this para is not urgently needed. agree to align with OECD 203 for para 24 regarding chemical analysis frequency.

13. para 23 - In this paragraph a guidance could be made on when a semi-static or even flow-through approach could be employed.

Rational: A possibility of performing this test under semi-static or flow-through conditions is mentioned only once, in paragraph 14 but no further guidance is given.

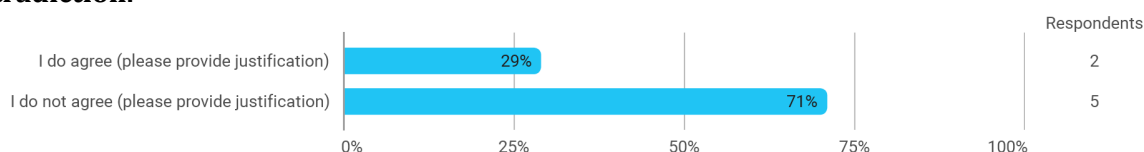


Disagreement justifications

- Flow through is not always possible, even when needed
- See section 14.
- Disagree, refer to OECD GD 23
- See section 14.
- Not relevant. This is TG23 guidance

14. para 24 - “limit test”: Suggested update: replace the phrase “in order to demonstrate that the EC50 is greater than this concentration” with “in order to demonstrate that the test substance has no toxic effects at this concentration”

Rational: The same paragraph states that “If the percentage of immobilisation exceeds 10% at the end of the test, a full study should be conducted”, therefore say that the objective of the limit test is to demonstrate that the effect is < 50% seems a contradiction.



Agreement justifications:

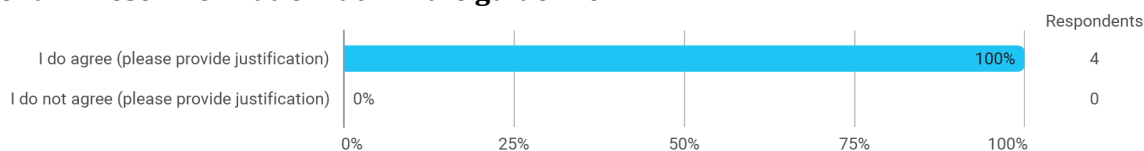
- correct as suggested

Disagreement justifications

- The parameter to be determined in an acute daphnid test is the EC50, not the NOEC. If it would be the NOEC, yes, adaptation of the sentence would be necessary. But this is not the case.
- this guideline is meant to provide information on EC50.
- Disagree, this is too restrictive. The main endpoint of the study is the EC50. A limit test is acceptable if some effects (immobilization or sublethal effects) are seen at the tested concentration but remain below the EC50. In addition: we cannot test higher when the limit test was performed at the solubility limit.
- The parameter to be determined in an acute daphnid test is the EC50, not the NOEC. If it would be the NOEC, yes, adaptation of the sentence would be necessary. But this is not the case.
- Disagree. Keep wording aligned with Threshold approach

15. para 25 - Guidance on most appropriate models should be given. Also, a guidance on how to treat concentrations strongly deviating from dose-response is necessary.

Rational: These information lack in the guideline

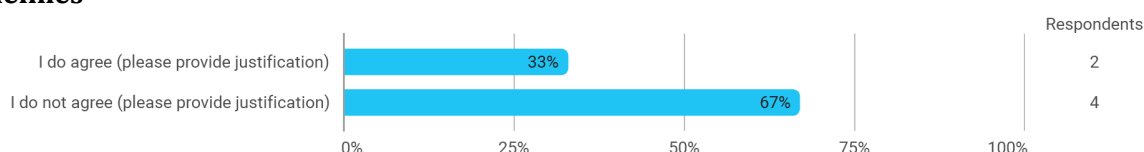


Agreement justifications:

- ▶ it rarely happens that daphnia (and also other organisms) do not react in a dose response, but if, then it is unknown how to continue
- ▶ agree
- ▶ May be useful, as long as it is not restrictive

16. Data and reporting: Add the following text from Para. 31 of OECD TG 203: It is recommended that results should be calculated using the measured concentrations of the test chemical. If the deviation from the nominal concentrations is smaller than 20%, results may also be based on the nominal concentrations. It should be noted that it is often useful to have both measured and nominal effect concentrations quoted, see Guidance Document No. 23 (OECD, 2019).

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines

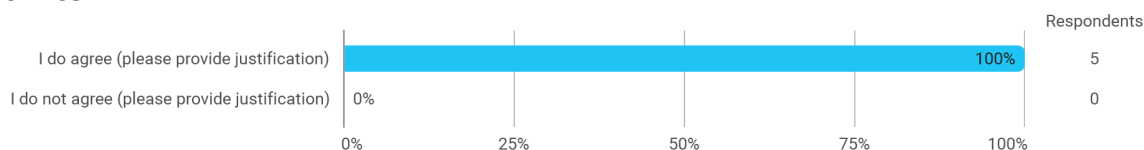


Disagreement justifications

- ▶ reporting endpoints on more than one concentration range and additionally for each active, which is often required, makes a report unclear, but if required by authorities I do not know) should be done, but maybe restricted to some parts of the report (e.g. only in results section)
- ▶ See para 23.
- ▶ The information given is sufficient. Different regulatory frameworks use different approaches for the calculations. Harmonization would be appreciated but might not be the right way to do it via the OECD technical guidelines.

17. Align Annex 3 with the similar Annex 3 from OECD TG 203.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines



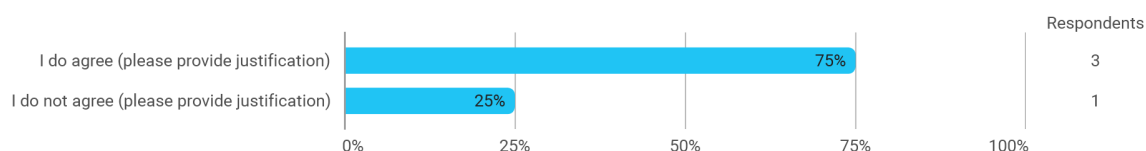
Agreement justifications:

- ▶ more detailed
- ▶ Align with other TGs

- agree
- Agreed In addition, the mistakes for some of the medium ingredients should be corrected and aligned with OECD TG 235

18. Definition of immobilisation needs to be more specific. Perhaps the word "within" should be replaced with "for at least" in "Those animals that are not able to swim within 15 seconds, after gentle agitation of the test vessel are considered to be immobilised (even if they can still move their antennae)."

Rational: There are people interpreting it in a way that the daphnids need to swim the whole 15 s period, while others will be considering a daphnid swimming once up and down within 15 s after agitation as a mobile daphnid.



Agreement justifications:

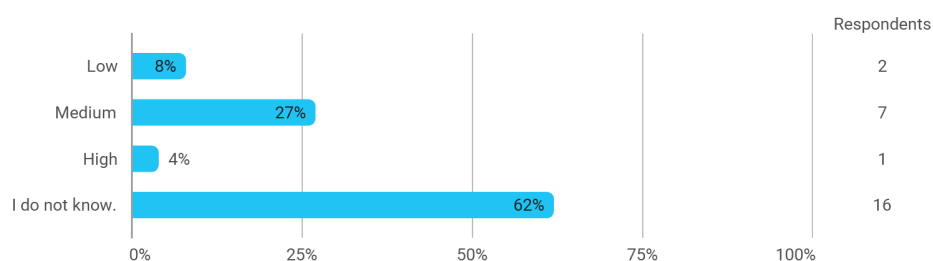
- different wording as suggested would change meaning, but could indeed be more specified
- definition is not clear

Disagreement justifications

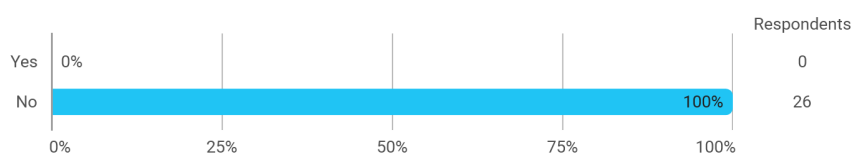
- Disagree. this wording has never been a problem for the use of this guideline.

C.12 Test No. 209: Activated Sludge, Respiration Inhibition Test (Carbon and Ammonium Oxidation)

Relevance of revision



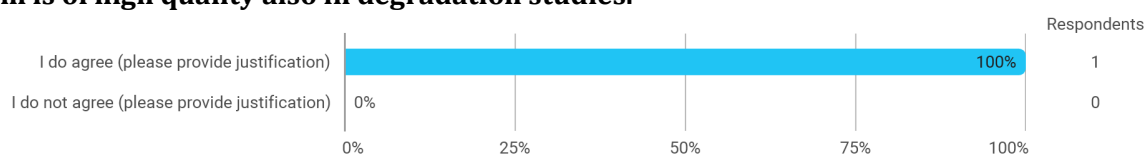
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

3. para 11- The blank controls (without the test substance or reference substance) oxygen uptake rate should not be less than 15 mg oxygen per one gram of activated

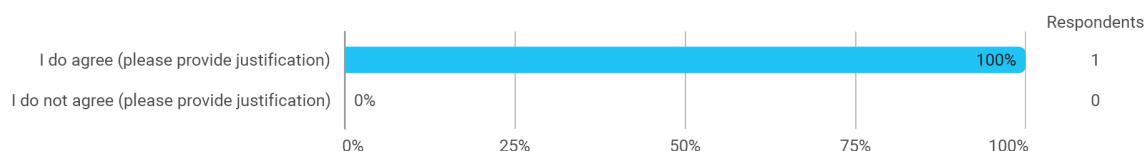
sludge (dry weight of suspended solids) in an hour. If the rate is lower, the test should be repeated with washed activated sludge or with the sludge from another source. Rational: This is a completely new validity criterion when compared to the version before 2010. From experience we can say that the blank control activity has no relevant influence on the EC values of the reference compound 3,5-DCP. Additionally, the guideline it selves states that the quality of inoculum varies immense between different STPs. So why a threshold value for background activity. From experience we know that – even after washing and feeding overnight – an oxygen uptake rate of 20 mg O₂/(g*h) is hardly to achieve with inoculum from STPs in northern Europe. In only approx. 50% of the tests, the threshold value is surpassed. However, the effects for the reference compound and all test substances so far were comparable in runs surpassing the validity criterion and not. If there is really a scientific reason for requiring a threshold value for basic activity, it should be lowered down at least to 15 mg O₂/(g*h). This value was achieved in most of the cases (however, only after washing and feeding over night!). By the way, our source is no industrial STP. The inoculum is of high quality also in degradation studies.



4. para 41 - The test should be carried out using a range of concentrations deduced from the preliminary test. In order to obtain both a NOEC and an ECx (e.g., EC50), six controls and five treatment concentrations in a geometric series with four replicates are in most cases recommended. In order to obtain both an ECx (e.g., EC50) only, four controls and five treatment concentrations in a geometric series with three replicates are in most cases recommended.

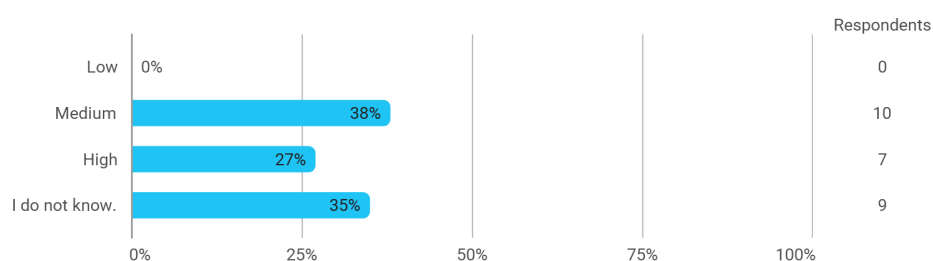
Rational: OECD 216 and 217 (Soil-Microorganisms) only require 3 replicates per treatment level. Even when no NOEC should be calculated, a statistical comparison of the treatment levels should be applied. At least for the EC calculation three replicates with 5 treatment levels are sufficient. Soil microorganisms and sludge microorganisms act comparable. Also, we are talking about thousands of individuals per replicate. However, even in chronic test designs with batches of animals or plants, four replicates are sufficient for NOEC and EC calculation! And in these test systems, higher variances within the replicates can be expected! For EC approaches only, the number of replicates for blank controls and treatment concentrations also can be reduced since the number of treatment levels is more relevant then the number of replicates. Since the guideline was adapted to general ecotox approaches, it is really a challenge to conduct this test design. And from experience we can say that neither the EC50, nor the EC10 is more representative after increasing the number of replicates! In contrast, the delay of the test run duration increases the variances. The inoculum biomass and activity in the stock increases in the duration of the start sequence of a sequential test design (at least after washing and feeding overnight). The respiration rate in the end blank control batch is always higher than in the start control batch! This could be due an increase of biomass in the course of the start sequence. However, it is technically not possible to measure (and adapt) the dry mass content before each application step. Therefore, for a representative result it is more important to run a

test in the shortest duration possible, than to blow up the test design as much as possible.

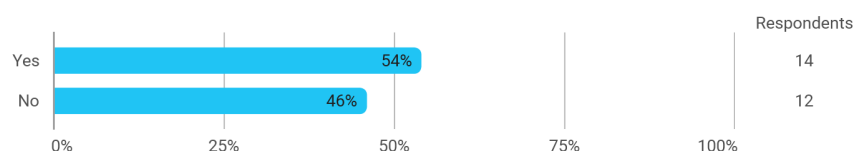


C.13 Test No. 211: Daphnia magna Reproduction Test

Relevance of revision



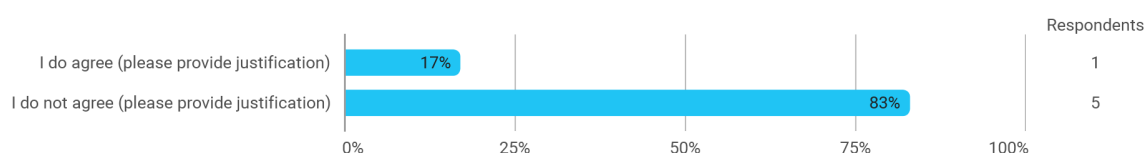
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. Add behavioural endpoints at the end of the test

Rational: Behavioural endpoints can be sensitive and highlight different information to standard endpoints of growth, reproduction and mortality. These endpoints can be measured during or at the end of standard OECD tests.

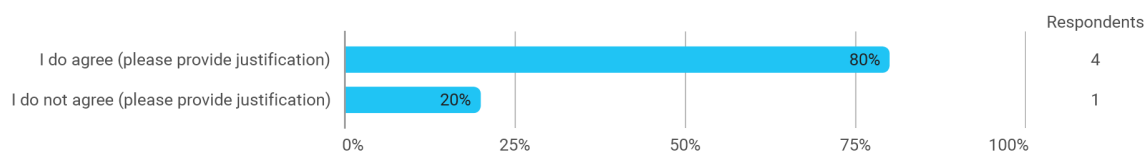


Disagreement justifications

- ▶ difficult to implement. where different behaviour like slow moving is seen also the number of offspring is usually affected
- ▶ Very subjective and difficult to interpret results
- ▶ Behaviour is recorded and any sublethal effect is described so, I don't think that additional endpoints are appropriate.
- ▶ Disagree. The test guideline has been validated and standardizes with a clear focus on deriving a NOEC and LOEC based on reproduction.
- ▶ Behavioural endpoints have not been validated

2. Include a section on initial considerations in line with the latest update of OECD TG 203 (18 June 2019).

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

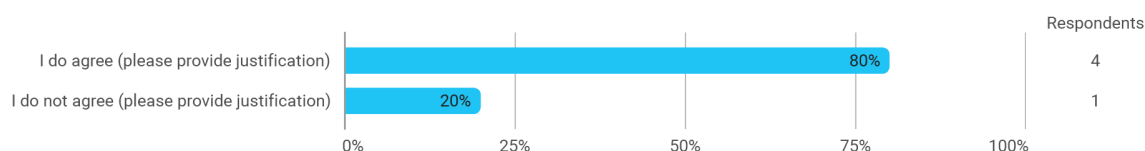
- may be helpful for guidance
- Align with other TGs
- Agree

Disagreement justifications

- Not relevant

3. para. 9. Add the following text from OECD TG 203: When testing difficult test chemicals, Guidance Document No. 23 (OECD, 2019) should be consulted and the study design modified appropriately. For volatile and other difficult test chemicals, further specific measures should be taken (OECD, 2018). Any silicone materials that were in contact with the test solution(s) should be preferably discarded and not reused for subsequent tests with different test chemicals.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.

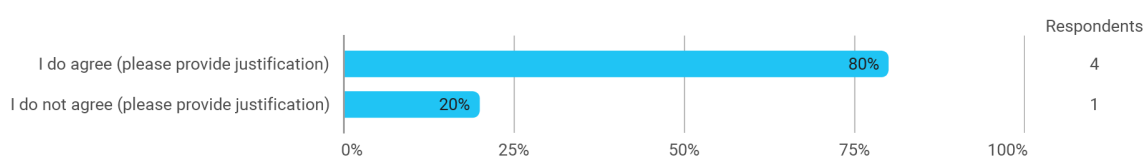


Agreement justifications:

- may be helpful
- Align with other TGs
- Agree

Disagreement justifications

- The reference to OECD TG 23 can be included in the future but is currently no reason for re-opening this guideline.
- 4. **para. 15: align the wording with OECD TG 203. OECD TG 211 states that measuring TOC/DOC is only recommended while TG 203 states it should be measured. It is proposed to use the following wording from OECD TG 203: If natural water is used, DOC or TOC and nitrate-content (NO₃) should be measured once prior to the test. Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.**

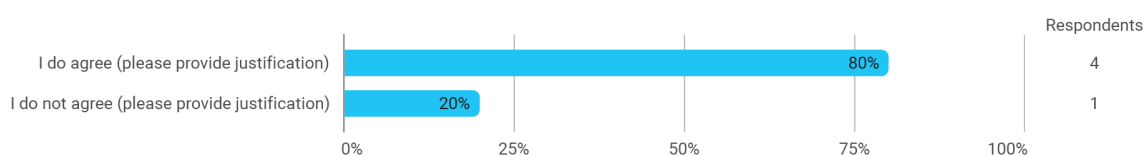


Agreement justifications:

- NO₃ is important
- Align with other TGs
- Agree

Disagreement justifications

- CROs are doing periodic checks of the DOC/TOC and nitrate but not before each test. Thus, this could be added but is not urgently needed.
- 5. **para. 18-19: to be replaced by Para. 17 from OECD TG 203. The following text from Para. 23 of OECD TG 203 should also be included: Low toxicity solvents only (i.e. acetone, ethanol, methanol, tertiary-butyl alcohol, acetonitrile, dimethyl formamide, dimethyl sulfoxide, and triethylene glycol) as recommended in Guidance Document No. 23 (OECD, 2019) should be used whilst solvents of unknown toxicity should not be used. It should be noted that in spite of the low toxicity for fish, dimethyl formamide and dimethyl sulfoxide should be avoided where possible on human health and safety grounds. Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.**



Agreement justifications:

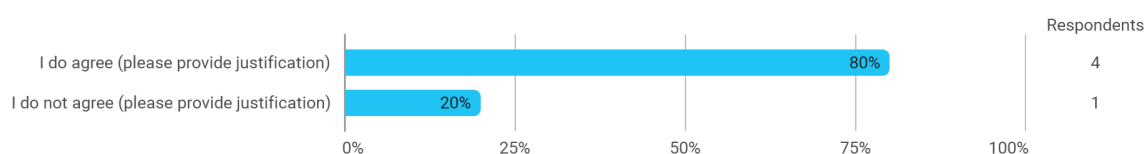
- ▶ can be replaced, but exist in OECD 211 in a similar way
- ▶ Align with other TGs
- ▶ Agree

Disagreement justifications

- ▶ DMF is a routinely used solvent in aquatic testing. It has been shown to be more suitable in most cases as e.g. acetone or ethanol. In general, untested substances, as routinely tested in ecotox labs, have to be handled as if they were CMR compounds. Thus at least the safety measures in the labs should always be appropriate. Based on the fact, that DMF was predominantly used by CRO's most of the historical data being available for possible effects of solvents, relate to DMF. The currently ongoing project investigating whether always a solvent control and a control is needed when a solvent is used in a study based their evaluations as well on a high number of studies which have been performed with DMF. The hope is, that this evaluation will allow to skip the negative control and by that reduce animal numbers in vertebrate tests. If we lose DMF that might have a negative impact on this decision.

6. Add Para. 18 from OECD TG 203.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

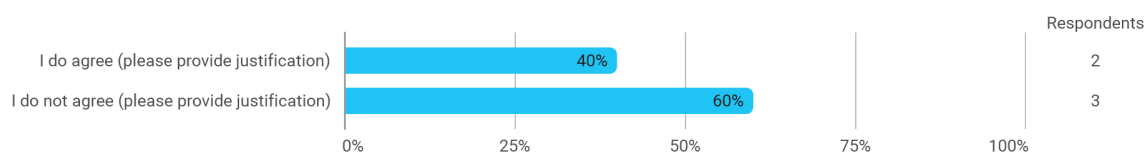
- ▶ no clear guidance with the pH adjustment is given now.
- ▶ Align with other TGs
- ▶ Agree

Disagreement justifications

- ▶ Not needed.

7. Frequency of analytical determinations and measurements (Para. 45-50): align with Para 24-25 of OECD TG 203.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

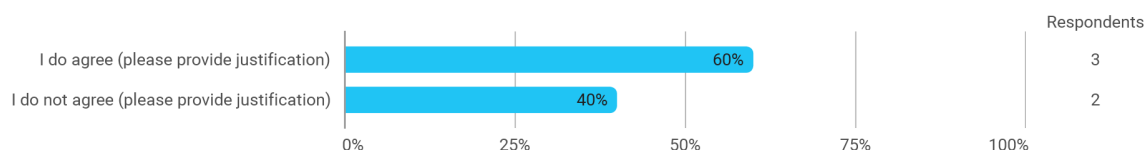
- Align with other TGs

Disagreement justifications

- Measurement of all water renewals would be a lot extra work.
- to add paragraph 24, 25 is not entirely relevant, daily measurements of pH and DO and for every vessel are not necessary and do not add to the results. Measurements from pooled media should be adequate
- disagree to align with paragraph 25 about the frequency of measurement of pH, O₂, salinity, etc. It is not necessary to measure them daily in all test vessels. agree to align with OECD 203 for para 24 regarding chemical analysis frequency.

- 8. Treatment of results: Add the following text from Para. 31 of OECD TG 203: It is recommended that results should be calculated using the measured concentrations of the test chemical. If the deviation from the nominal concentrations is smaller than 20%, results may also be based on the nominal concentrations. It should be noted that it is often useful to have both measured and nominal effect concentrations quoted, see Guidance Document No. 23 (OECD, 2019).**

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

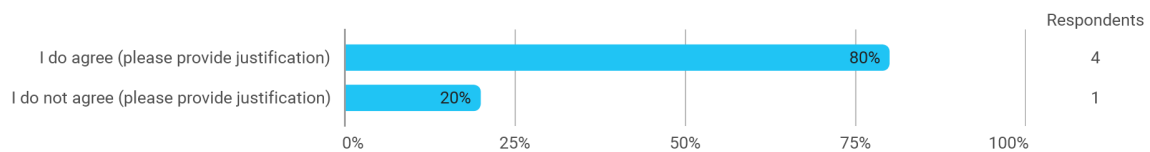
- Align with other TGs
- We don't agree with presenting both based on nominal and mean measured, but we agree with the need of clear guidance for unstable substances.

Disagreement justifications

- if reporting of nominal and measurement is required by authorities then useful otherwise not necessary as it is additional work and makes a report unclear
- Disagree. The information given is sufficient. Different regulatory frameworks use different approaches for the calculations. Harmonization would be appreciated but might not be the right way to do it via the OECD technical guidelines.

9. Include the Annex 3 from OECD TG 203 or at least include consistent information related to the content in particulate matter, TOC and COD in the description what makes an acceptable dilution/test water.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

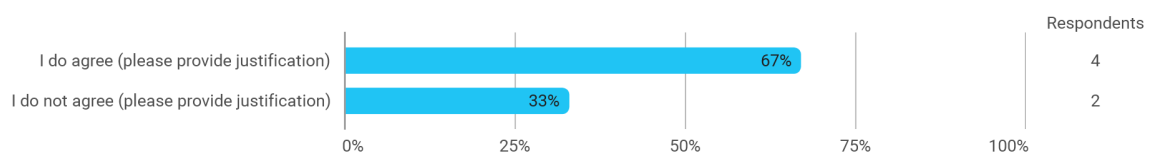
- ▶ useful information
- ▶ Align with other TGs
- ▶ agree

Disagreement justifications

- ▶ Disagree. Some information is already given with respect to water hardness. In general, it is not possible to copy from OECD TG 203 as Daphnia demands are different. Consider Annex 2.
- ▶ Not clear what this comment means.

10. para 6- More guidance should be given on how the results of an acute daphnia test might be useful in selecting an appropriate range of test concentrations.

Rational: At this moment this paragraph is vague.



Agreement justifications:

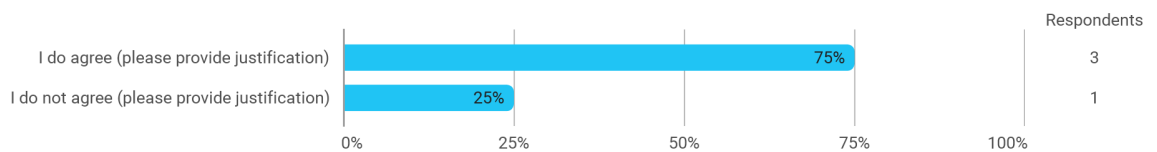
- ▶ would be helpful
- ▶ this would be very useful and reduced the need for 211 range finding
- ▶ agree

Disagreement justifications

- ▶ Disagree
- ▶ Could be helpful, but is not needed. This is what range finding tests are for

11. para 17 - "The pH should be within the range 6 - 9, and normally it should not vary by more than 1.5 units in any one test"

Rational: It should be clear what is meant with the variation of 1.5 units; the difference between the control and test groups or the change during the exposure. Both can be difficult to archive especially that it is advised not to adjust pH.



Agreement justifications:

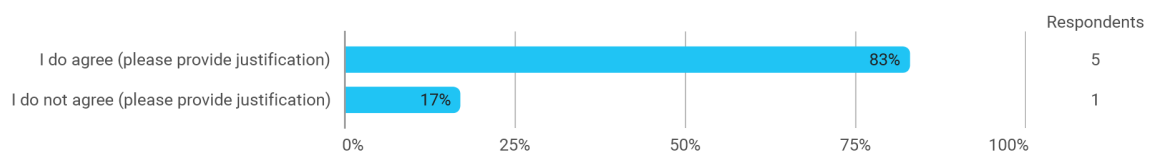
- It is not clear what should be changed, but the sentence is not really clear written in the guideline.
- Align with other TGs

Disagreement justifications

- Not clear what this comment means.

12. para 30 - "at an intensity not exceeding 15-20 $\mu\text{E}/\text{m}^2/\text{s}$ ". It seems that it should be either "between" or "not exceeding 20 $\mu\text{E}/\text{m}^2/\text{s}$ ".

Rational: It's confusing at this moment.



Agreement justifications:

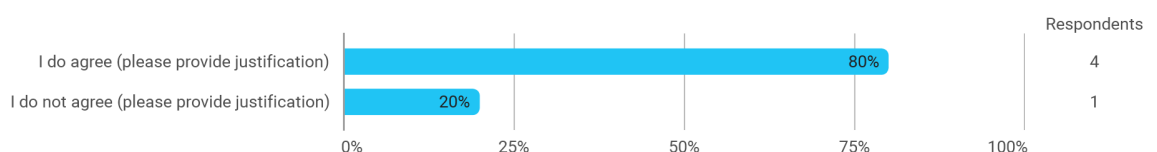
- definitely need review. This sentence always end in discussions with monitors.
- wording/clarity
- agree
- Yes, needs better wording

Disagreement justifications

- Disagree, not needed.

13. para 38 - The text of this paragraph should be included in validity criteria.

Rational: In case variability in the control exceeds 25% statistical analysis is pointless unless reproduction is completely inhibited at test concentrations.



Agreement justifications:

- but only for surviving adults.

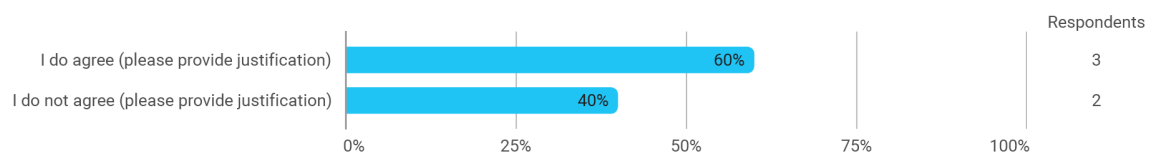
- ▶ some of our clients have pulled this out and added it as a validity criteria. It defines a higher quality study
- ▶ agree

Disagreement justifications

- ▶ Disagree. Changes of validity criteria cannot be based on single comments. Here it would be needed to investigate underlying data to check whether a change is reasonable and needed! Eventually a new ring test would be needed.

14. para 39 - A recommendation could be made to perform a pre-test assessing stability of the test item in medium for a period of 24, 48 and 72 h.

Rational: This could provide solid base for setting up the renewal frequency.



Agreement justifications:

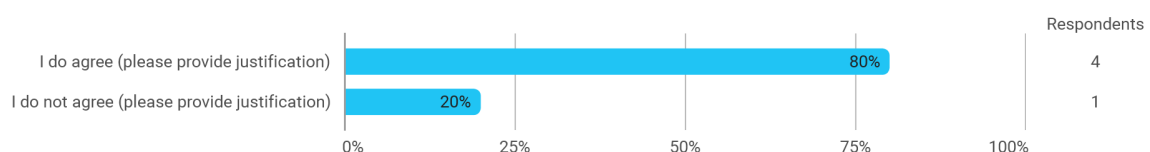
- ▶ could be included, but recommendations are often interpreted as a "must have"
- ▶ agree

Disagreement justifications

- ▶ a suggestion but not a recommendation. Generally we have acute data that can support this.
- ▶ Disagree. This can be done by performers already. A recommendation in a guideline is already rather strong wording and would result in additional not necessary efforts. In most cases information on stability is available before running a chronic Daphnia study. In addition, Daphnia repro studies are performed mostly under semi-static or flow through conditions to allow appropriate water quality and stability of the test item over time.

15. para 44 - What if growth is more sensitive to a given item than reproduction, i.e., NOEC growth < lowest concentration tested but NOEC reproduction ≥ lowest concentration. Need the study be repeated to refine the NOEC for growth? Clarification should be added.

Rational: Although the assessment of effects on reproduction is the primary goal of this study, addition of clarification what to do in case of above mentioned situation might prevent unnecessary confusion/discussions.



Agreement justifications:

- ▶ may help
- ▶ agree

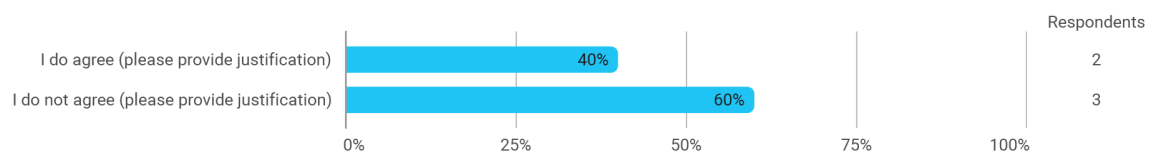
- More explicit guidance on selecting the driving endpoint may be useful

Disagreement justifications

- Disagree, no further information needed.

16. para 51 - The parameter offspring of dead parents should be included in the statistical evaluation.

Rational: Statistical evaluation not entirely clear if significant effects occur in mortality but no effects occur in reproduction of the dead parents. In this case the effect for reproduction of introduced parents is always significant, but this doesn't indicate the real effect of offspring.



Agreement justifications:

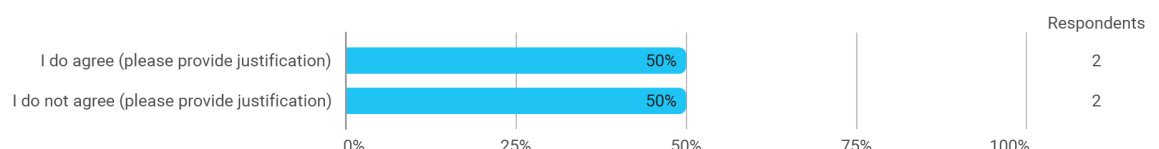
- it is already in the TG when there is a dose response for mortality see paragraph 4

Disagreement justifications

- Do not understand
- This will bring no additional value
- Disagree, no need for a change. Dead parental animals can relate to e.g. handling of animals in a semi static test and thus the observed mortality is not related to the test item. Current version of TG therefore is sufficient. The ecologically most relevant response variable is the total number of living offspring per parent animal which does not die accidentally or inadvertently. The evaluation including offspring of dead parental animals is not excluded by the guideline and can voluntarily be included already.

17. para 56 - A sentence could be added like “in cases a stimulation rather than reduction of reproduction or growth is observed at all concentrations when compared with the control, no statistical analysis is required.”

Rational: This seems to be quite obvious but can prevent unnecessary confusions/discussion.



Agreement justifications:

- agree

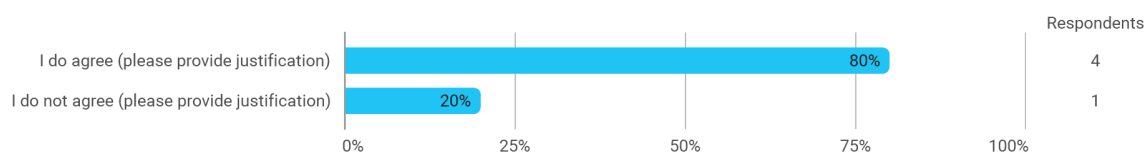
Disagreement justifications

- rare case and a statistical analyses does not hurt

- No need for a change. But it is totally correct that an EC20 calculation is not always reasonable.

18. It would be beneficial to include a decision tree describing the statistical procedure.

Rational: It is easier to follow than narrative description.



Agreement justifications:

- would be helpful, but may also be the right place for in the OECD 54 (if not already available)
- this would be helpful for all TGs
- agree - but only to be used as a recommendation and not as a strict guidance which could bring limitations and major changes required to historical studies

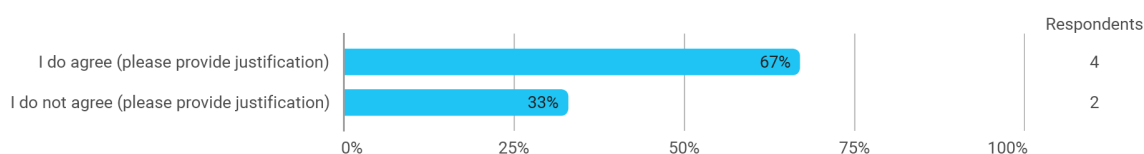
Disagreement justifications

- Support for Statistical procedures is given in other OECD documents and can be referred to.

19. Handling of volatile substances

Although daily renewal for very unstable substances does not cause any problems (besides additional work), but for an unknown reason, it is not possible to perform in closed vessels with highly volatile substances (very high mortality in the control, personal experience). Employing renewal every 48 hours solves this problem. This should be further investigated and if necessary, included in the guideline. Many sponsors require daily renewal and are convinced that the lab not the setup are responsible for the failure.

Rational: This is a personal experience with multiple studies with very volatile substances.



Agreement justifications:

- daily renewal can indeed cause problems
- this would support the testing of volatiles and expectations/validity
- agree

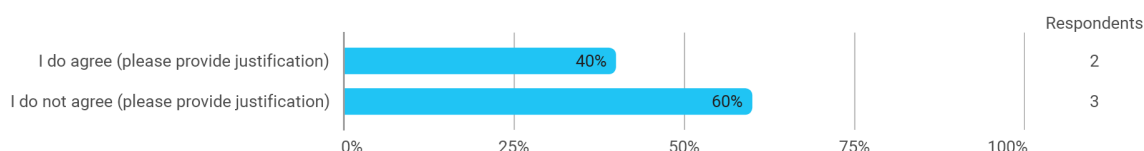
Disagreement justifications

- For the handling of volatile substances the first step should be to consider OECD TG 23. Daily transfer of Daphnia is already problematic as this procedure is stressing Daphnia. How to best use closed vessel systems needs to be investigated further. This procedure has Pros and Cons and can not just be changed or added to the guideline without further standardization.

Summarizing it could be useful to add clarification on this point in the future version of this guideline. But nothing which can be changed on short notice.

- This is TG23 related

20. The methodology as described in the current guideline is not suitable to test for effects of virtually insoluble particulate materials (e.g. nano-forms of substances, polymers). Adaption of the guidelines methodology with latest scientific findings is indicated so that valid tests with particulate materials can be performed.



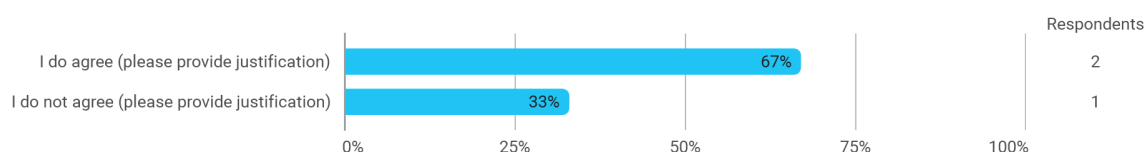
Agreement justifications:

- how should this be possible?

Disagreement justifications

- this should be handled in a separate guideline
- The test was standardized and validated for dissolved test items. For difficult substances please refer to the OECD Technical Guidance on difficult substances (OECD GD 23).
- Not relevant for this guideline. This discussion is covered in TG23

21. paras 19,37, 59, 60 - Chapters should be adapted such, that the use of auxiliary agents such as chemical dispersants or emulsifying auxiliary agents is not not advocated, because of the potential for physical or chemical interactions influencing the apparent toxicity of the test chemical. Appropriate solvents used in conjunction with physical mixing is allowed.



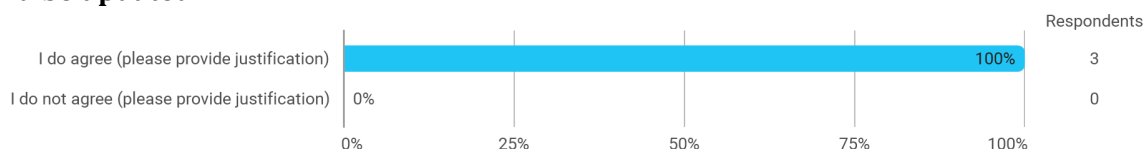
Agreement justifications:

- dispersants should be avoided, but other solution is needed then

Disagreement justifications

- this is already clear enough in the guideline. In paragraph 19 there is a sentence saying: A dilution water control with adequate replicates and, if unavoidable, a solvent control with adequate replicates should be run in addition to the test concentrations.

22. para 18 and Literature - reference (13) and corresponding text in paragraph 18 should be updated

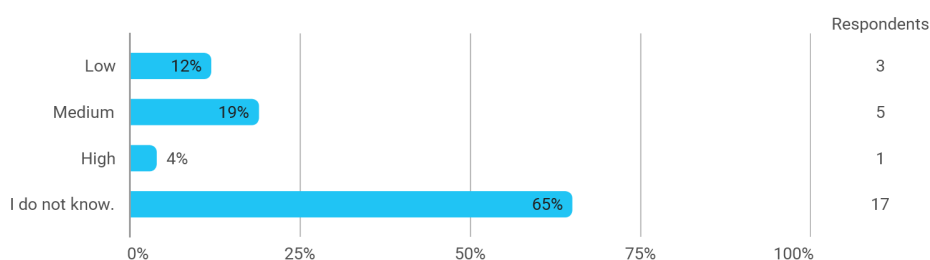


Agreement justifications:

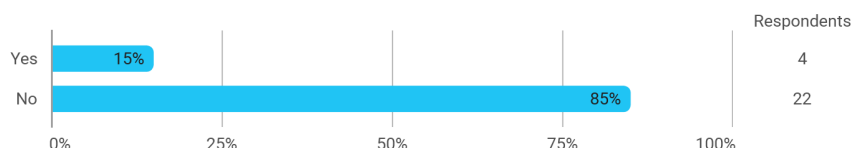
- out of date
- agree

C.14 Test No. 243: *Lymnaea stagnalis* Reproduction Test

Relevance of revision

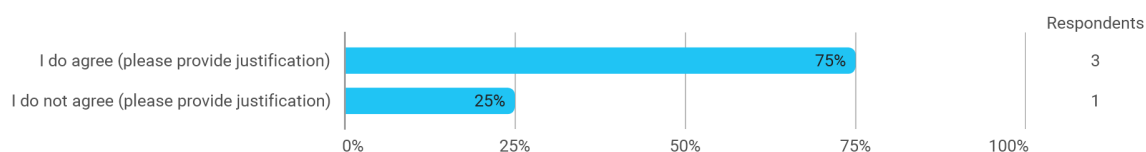


Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. para 10 - For a test to be valid, the following criteria should be met: the mean control mortality (accounting for all control replicates) should not exceed 20% at the end of the test; the mean cumulated individual fecundity in the controls at the end of the test should be at least four egg-clutches per individual over the 28 day test duration (see § 49); the dissolved oxygen content should be at least 60% of the air saturation value in both control and exposure groups throughout the test and, water mean temperature should be $20 \pm 1^\circ\text{C}$ throughout the test in both control and exposure groups. Transitory deviation (for 1 to 2 days) from this mean value can occasionally occur but should not be more than $\pm 2^\circ\text{C}$.
Rational: The original wording in the second bullet point ("the mean cumulated individual fecundity in the controls at the end of the test should be at least four egg-clutches per individual-day (see § 49);") has been mis-interpreted in the way that each individual is expected to produce at least 4 egg-clutches per day. This is far beyond the biological capacity of the snails. The new wording is more traceable for the users of the guideline.



Agreement justifications:

- I think this is that same as suggestion 2
- I agree on how Paragraph 10 has been worded

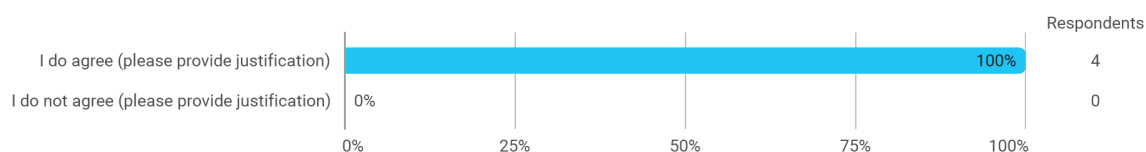
Disagreement justifications

- Not clear where the error is

2. **para 10 - The second validity criteria is believed to be incorrect/poorly worded. 'The mean cumulated individual fecundity in the controls at the end of the test should be at least four egg-clutches per individual-day' It should read similar to 'the mean number of clutches per individual in the controls should be at least 4 clutches at the end of the test'.**

3. **Rational: The points below give evidence and justification for changing the validity criteria. The reproductive output (clutches/ individual day) for the control our studies has produced similar results to the ring test laboratories which produced approximately 0.2 – 0.3 clutches/individual day (Figure 2, Ref 1). Therefore, we believe the control fecundity in this study to be sufficient. It should also be noted that the snails would have had to produce 10 x more clutches to meet the validity criteria given in the OECD 243 test guidance document which is not plausible. The ring test paper (Ref 1) states that the validity criteria for control snail fecundity over a 56-day exposure period should be >8 egg masses per snail over the test period which equates to 4 clutches over a 28-day exposure period.**

Ducrot, V., Askem, C., Azam, D., Brettschneider, D., Brown, R., Charles, S., Coke, M., Collinet, M., Delignette-Muller, M-L., Forfait-Dubuc, C., Holbech, H., Hutchinson, T., Jach, A., Kinnberg, K.L., Lacoste, C., Le Page, G., Matthiessen, P., Oehlmann, J., Rice, L., Roberts, E., Ruppert, K., Elphinstone Davis, J., Veauvy, C., Weltje, L., Wortham, R., Lagadic, L. (2014), Developmnet and Validation of an OECD reproductive toxicity test guideline with the pond snail *Lymnaea stagnalis* (Mollusca, Gastropoda). *Regulatory Toxicity and Pharmacology*, 70, 605-614.

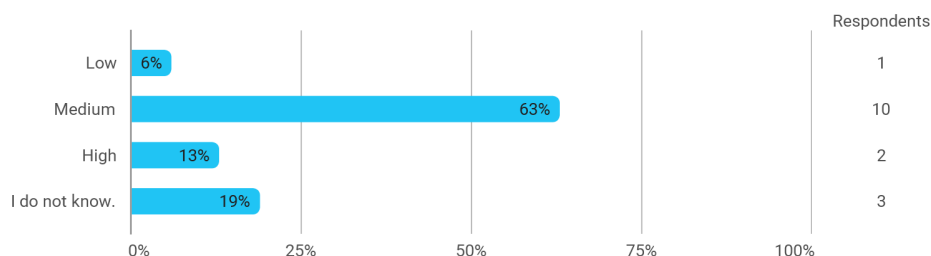


Agreement justifications:

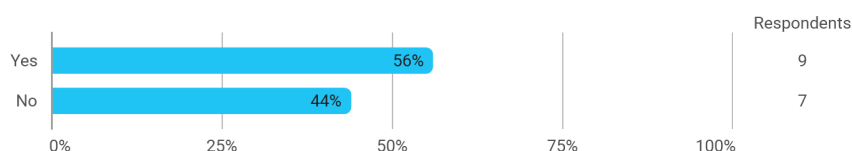
- We have extensive experience at our lab and have found this to be true
- I agree to the suggested changes

C.15 Test No. 218: Sediment-Water Chironomid Toxicity Using Spiked Sediment

Relevance of revision



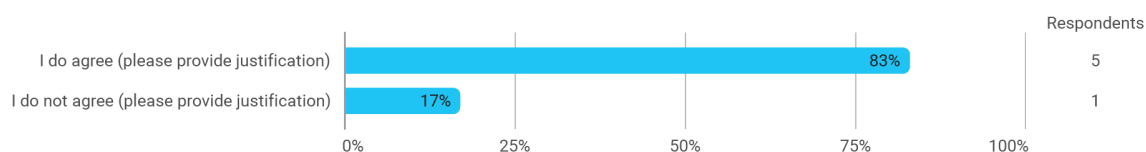
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. para 10 - The validity criterion in the first bullet point of §10 is not clear. It says: “For the test to be valid the following conditions apply: - the emergence in the controls must be at least 70% at the end of the test (1)(6); ...” Does this mean that each individual replicate should have at least 70% emergence, or does it mean that the mean emergence rate across replicates for the control group should be at least 70%? This validity criterion should be clarified.

Rational: A similar question, but for TG 208, was raised by the OECD secretariat to the NC for the ENV in an e-mail 2021-05-05. In this e-mail it was mentioned that OECD regularly receive similar questions also on other tests where the wording is also not straightforward. We have noted that the wording for TG 218 could be clarified considering this. (Based on the feedback received for the TG 208-question, OECD made the following conclusion (e-mail 2021-06-18): “From feedback received it appears that the validity criterion in TG 208 is mostly interpreted as applying to the mean of all replicates in the control group, rather than to each replicate (in TG 208 a replicate can contain one or two large seed(s) only, depending on the size of the seed). For now, I will respond in that sense to the enquiry received. However, your feedback also indicates that it might be welcome to report the variance across replicates. The TG 208 is generally not very prescriptive for the test design (e.g. no minimum number of replicates), nor does the test report explicitly requires individual replicate data it seems. Like for many TGs drafted 10-15 years ago or more, the level of prescription is not high. If the TG 208 was to be written today, more precision would be expected, based on experimental validation work. I also take from feedback received that it may not be appropriate to have the same way of interpreting the validity criterion in all Test Guidelines, as test designs vary. Thank you and for now, I will leave this conversation here. If there was a need to re-open and clarify things further, a clear project proposal would be needed.”)



Agreement justifications:

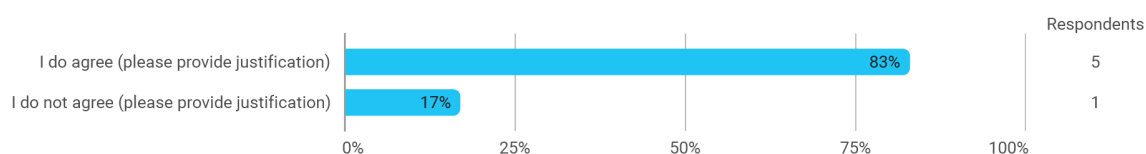
- ▶ should be specific
- ▶ Agree: Diverging interpretation between different regulators has been observed. Since this is about the validity of the study this is important to clarify. Clarification is needed. As far as remembered this was originally related to the mean of all replicates. the mean of all replicates.
- ▶ Additional clarify could be useful though it is the validity criterion is general focused on group, not replicate response.
- ▶ So far, we used the mean emergence rate across replicates

Disagreement justifications

- ▶ I think it is clear, that validity is the mean emergence rate (as in other test systems)

2. para 10 - In the second point, for the sentence “C. riparius and C. yoshimatsui emergence to adults from control vessels should occur between 12 and 23 days after their insertion into the vessels” it should be specified whether the emergency in the control group should begin between days 12 and 23 or even end (or other).

Rational: It is unclear whether the validity criterion is fulfilled in the event that some adults of the control group emerge after day 23. Moreover, for the first validity criterion the value of 70% is calculated at the end of the test and in OECD 233 (2010) at paragraph 10 it is specified that 85% should emerge in the period 12-23 days.



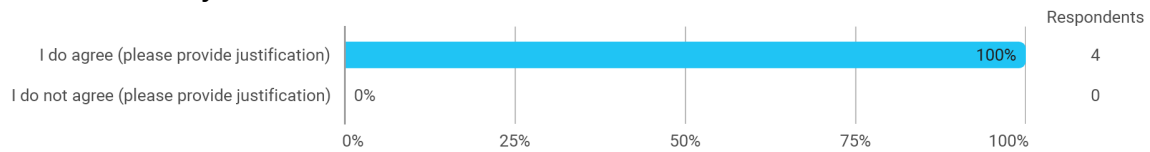
Agreement justifications:

- ▶ current definition is not clear
- ▶ I think it should be clarified
- ▶ Agree: Diverging interpretation between different regulators has been observed. Since this is about the validity of the study this is important to clarify
- ▶ Usually a few larvae hatch after day 23

Disagreement justifications

- ▶ Disagree. No concern. Additional clarity could be useful though I believe the current text is adequate.

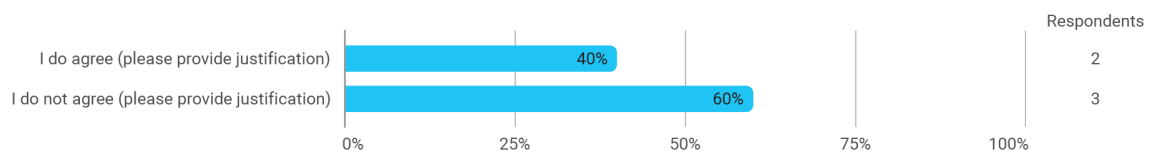
3. para 13 - Other alternatives for sphagnum peat should be included in the guidelines.
Rational: Sphagnum peat is not available in all countries and it's also not environmentally sustainable.



Agreement justifications:

- ▶ its difficult to source and environmentally bad to use
- ▶ Partly agree: if these alternatives are known to work
- ▶ The guideline should allow flexibility for the test facility in terms of animal husbandry in it optimizes performance of the test system.

4. para 16 - Food should be spiked with the test compound
Rational: The need of not adding naive food therefore adding food before spiking has been already addressed (EFSA Journal 2015;13(7):4176) and was also pointed out during the update of the REACH sediment risk assessment recommendations.



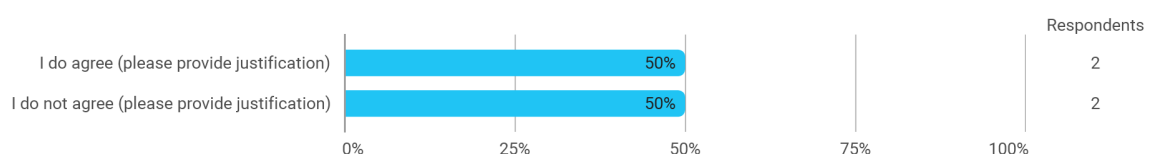
Agreement justifications:

- ▶ Having more detail about the types, rations, and approaches for providing food is always beneficial in guideline.

Disagreement justifications

- ▶ contaminated food is a different exposure scenario than spiked sediment
- ▶ why?? Clients/regulators may not want this. changes the study design.
- ▶ Strongly disagree: This would be another guideline then, evaluation of the practicality is needed + ring test This has never been subject of standardization or validation of the current guideline. The objective of the guideline is not to investigate the impact of feeding only. The effect of the test item via contact to the sediment and contact to overlying and porewater in combination with normal food uptake has been subject for the set-up of this guideline.

5. para 19 - a higher max. spacing factor should be allowed



Agreement justifications:

- ▶ higher spacing factor could empower statistical tests

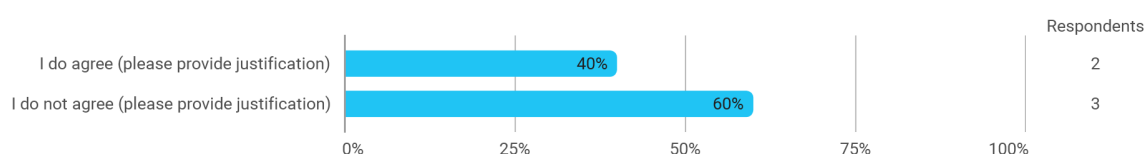
Disagreement justifications

- If the focus is an ECx value, then narrowing spacing may be needed than for a hypothesis NOEC approach. With that said, it would be good to understand the technical basis by which a spacing of 2 was selected.

- Not necessary

6. para 24 - The stabilisation period should be harmonised as much as possible, e.g. if using both OECD and US EPA/ASTM test protocols to avoid differences in exposure concentrations due to different ageing periods.

Rational: Guidance on stabilization/equilibration period and conditions similar to what is provided in other OECD guidelines are needed. If not possible to provide a unique protocol that serves all, maybe some indication for substances with different properties.

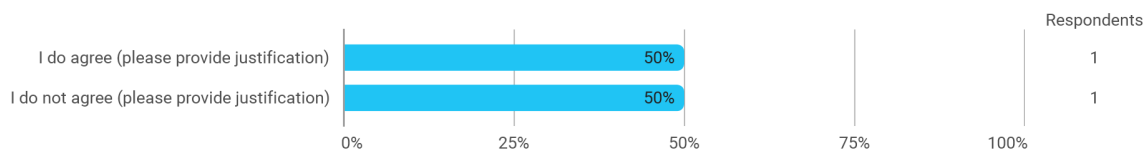


Agreement justifications:

- It would be nice to have more clarity on this

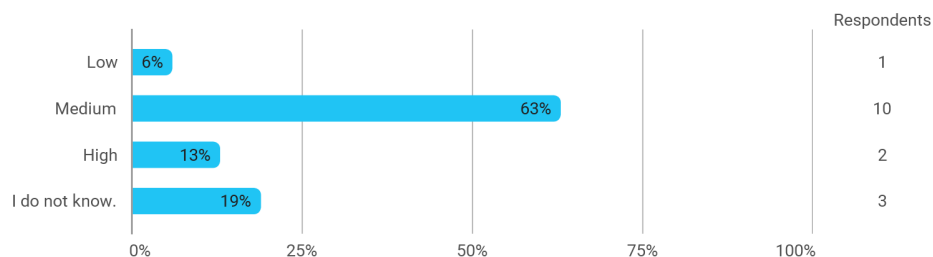
Disagreement justifications

- should be up to the labs experience
 - Disagree: The US guidelines and the OECD guidelines have different objectives (prospective vs retrospective assessment) and the design is different static vs semi-static test. Alignment is therefore not possible and not desired by the different frameworks. Regulators should then make sure to use the corresponding endpoints in an appropriate manner, rather than expecting alignment. The tests are performed for different purposes and can under these circumstances not be harmonized. A harmonization would first an alignment on whether a prospective or a retrospective sediment risk assessment approach is needed.
 - Currently, not only do approaches for stabilization differ between the US and OCED approach, but they also differ between testing facilities for the same guideline. I do not agree with harmonization, but would strongly support the inclusion of additional details concerns options for this approach and then allowing flexibility by situationally adapting to the study design based on the research question and properties of the compound. While not specifically mentioned, is there the possibility to eliminate the negative solvent control treatment if solvent is used as a transient carrier as would be consistent with the US EPA guidance. The particular way the solvent is used in a study could also impact equilibration/stabilization.
7. para 38 - Please refer to the recently provided Ctgb comments on (by UBA) proposed revisions of OECD 2018 and 2019 on CIRCABC, which are in line with the comments provide on OECD 233 above: <https://circabc.europa.eu/ui/group/0b40948d-7247-4819-bbf9-ecca3250d893/forum/topic/617aa6a7-4f06-4485-9275-809ab5d77cb8>
Rational: Ambiguity/ unclarity on what is required, particularly for fast degrading substances, which may potentially invalidate the test.

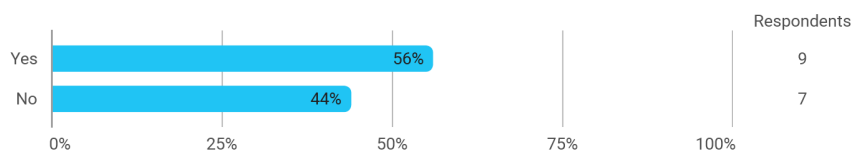


C.16 Test No. 219: Sediment-Water Chironomid Toxicity Using Spiked Water

Relevance of revision



Interest in being part of an OECD Expert Group

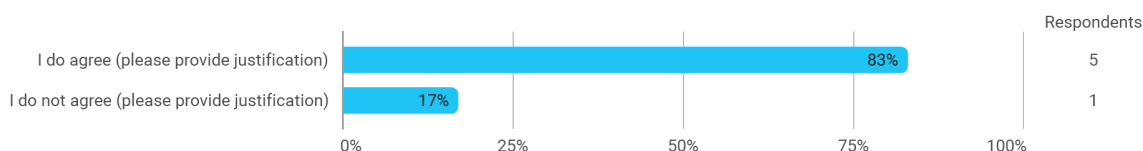


The following suggestions have been collected:

- para 10 - The validity criterion in the first bullet point of §10 is not clear. It says: “For the test to be valid the following conditions apply: - the emergence in the controls must be at least 70% at the end of the test. (1)(6); ...” Does this mean that each individual replicate should have at least 70% emergence, or does it mean that the mean emergence rate across replicates for the control group should be at least 70%? This validity criterion should be clarified.**

Rational: A similar question, but for TG 208, was raised by the OECD secretariat to the NC for the ENV in an e-mail 2021-05-05. In this e-mail it was mentioned that OECD regularly receive similar questions also on other tests where the wording is also not straightforward. We have noted that the wording for TG 219 also could be clarified considering this. (Based on the feedback received for the TG 208-question, OECD made the following conclusion (e-mail 2021-06-18): “From feedback received it appears that the validity criterion in TG 208 is mostly interpreted as applying to the mean of all replicates in the control group, rather than to each replicate (in TG 208 a replicate can contain one or two large seed(s) only, depending on the size of the seed). For now, I will respond in that sense to the enquiry received. However, your feedback also indicates that it might be welcome to report the variance across replicates. The TG 208 is generally not very prescriptive for the test design (e.g. no minimum number of replicates), nor does the test report explicitly requires individual replicate data it seems. Like for many TGs drafted 10-15 years ago or more, the level of prescription is not high. If the TG 208 was to be written today, more precision would be expected, based on experimental validation work. I also take from feedback received that it may not be appropriate to have the same way of interpreting the validity criterion in all Test Guidelines, as test designs vary. Thank you and for now, I will leave this

conversation here. If there was a need to re-open and clarify things further, a clear project proposal would be needed.”)

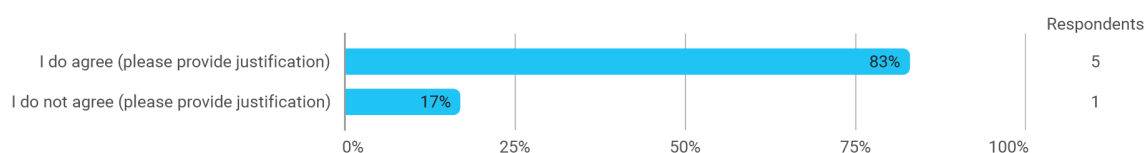


Agreement justifications:

- ▶ should be clarified
- ▶ Agree: Diverging interpretation between different regulators has been observed. Since this is about the validity of the study this is important to clarify Clarification is needed. As far as remembered this was originally related to the mean of all replicates. the mean of all replicates.
- ▶ Additional clarify could be useful though it is the validity criterion is general focused on group, not replicate response
- ▶ So far, we used the mean emergence rate across replicates

Disagreement justifications

- ▶ I think it is clear (mean emergence rate as in other systems)
- 2. para 10 - In the second point, for the sentence “C. riparius and C. yoshimatsui emergence to adults from control vessels should occur between 12 and 23 days after their insertion into the vessels” it should be specified whether the emergency in the control group should begin between days 12 and 23 or even end (or other). Rational: It is unclear whether the validity criterion is fulfilled in the event that some adults of the control group emerge after day 23. Moreover, for the first validity criterion the value of 70% is calculated at the end of the test and in OECD 233 (2010) at paragraph 10 it is specified that 85% should emerge in the period 12-23 days**



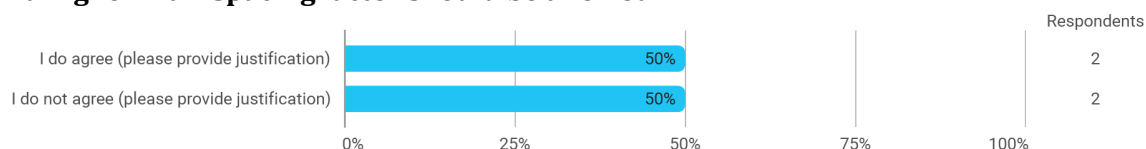
Agreement justifications:

- ▶ current description is not clear
- ▶ could be clarified
- ▶ Agree: Diverging interpretation between different regulators has been observed. Since this is about the validity of the study this is important to clarify
- ▶ Usually a few larvae hatch after day 23

Disagreement justifications

- ▶ No concern. Additional clarity could be useful though I believe the current text is adequate

3. para 19 - a higher max. spacing factor should be allowed



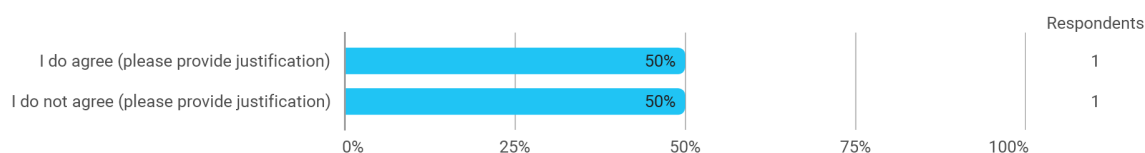
Agreement justifications:

- higher spacing factor could empower statistical tests

Disagreement justifications

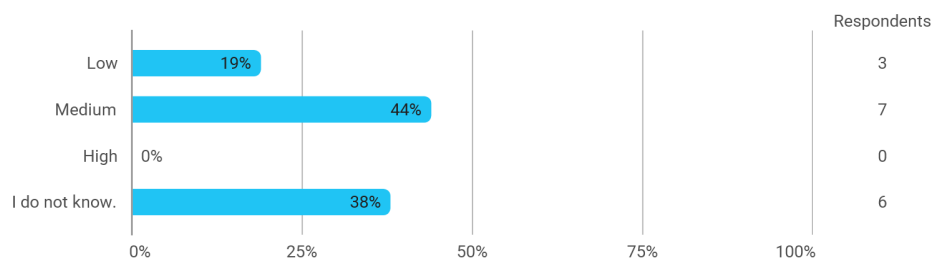
- If the focus is an ECx value, then narrowing spacing may be needed than for a hypothesis NOEC approach. With that said, it would be good to understand the technical basis by which a spacing of 2 was selected.
- Not necessary

4. para 38 - Please refer to the recently provided Ctgb comments on (by UBA) proposed revisions of OECD 2018 and 2019 on CIRCABC, which are in line with the comments provide on OECD 233 above: <https://circabc.europa.eu/ui/group/0b40948d-7247-4819-bbf9-ecca3250d893/forum/topic/617aa6a7-4f06-4485-9275-809ab5d77cb8>
Rational: Ambiguity/ unclarity on what is required, particularly for fast degrading substances, which may potentially invalidate the test.

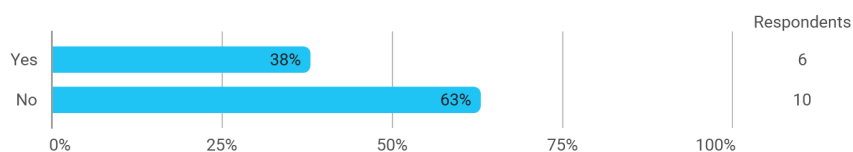


C.17 Test No. 225: Sediment-Water Lumbriculus Toxicity Test Using Spiked Sediment

Relevance of revision



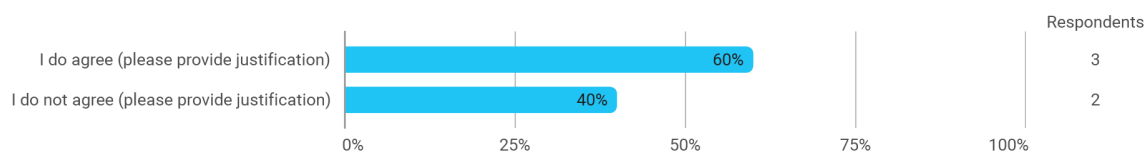
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. para 53 - Suggestion to refine text: 'In case of specific substance properties, e.g., if rapid degradation of the test substance is expected, the analytical schedule may be refined (e.g., more frequent sampling, analysis of more concentration levels) on the

basis of expert judgment. Samples may then be taken on intermediate sampling dates (e.g. on day seven after start of exposure), in both relevant compartments, and for all test concentrations in order to facilitate the calculation of a mass balance and of endpoints for risk assessment based on geometric mean measured concentrations. Rational: Ambiguity/ unclarity on what is required, particularly for fast degrading substances, which may potentially invalidate the test.



Agreement justifications:

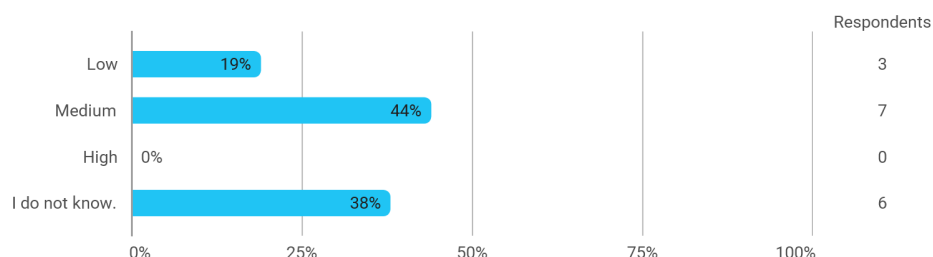
- good point to add
- Minimum requirements often not sufficient

Disagreement justifications

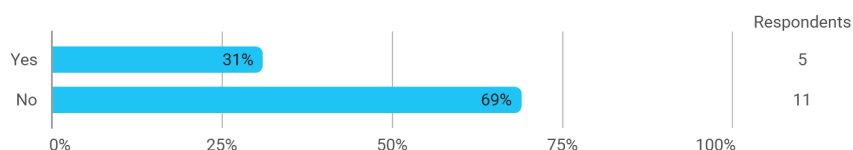
- Guideline says "at least", so you're always free to do more and adapt the sampling and evaluation to the needs of the test substance
- VERY STRONGLY DISAGREE. Focusing on mass balance during an ecotoxicity study adds substantial complexity and cost, but would provide minimal value for defining the regulatory hazard endpoint. E fate studies should be performed to better understand mass balance separately

C.18 Test No. 233: Sediment-Water Chironomid Life-Cycle Toxicity Test Using Spiked Water or Spiked Sediment

Relevance of revision

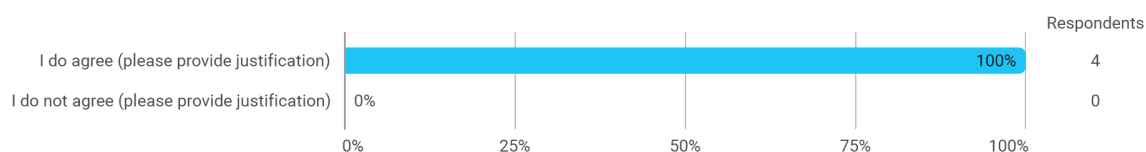


Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. para 9 - This sentence can potentially be removed: 'However, until a validated acute guideline is available a chronic test according to OECD 219 may be considered.'
Rational: Acute test guideline OECD 235 is available.

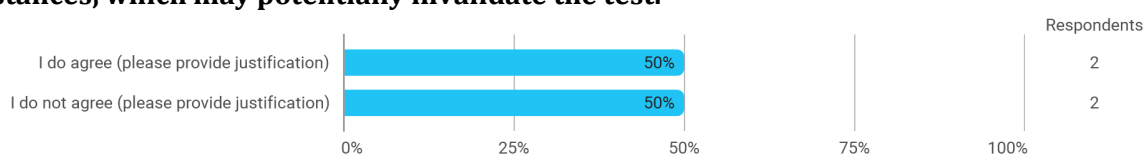


Agreement justifications:

- TG for acute test is available
- Agreed, but not urgent.
- agree

2. **paras 38, 39 and 22b - It is unclear from the current guideline, if for this study the tier 1 requirement of continuous exposure throughout the complete study applies as well and if therefore in case of fast degrading substances the exposure should also be expressed based on geometric mean measured concentrations, if it cannot be confirmed to lie between 80% and 120% of nominal throughout the test (in the whole system). For a tier 1 test this would however be required from a risk assessment perspective. This consequently means that the analytical determinations required in case of fast degrading substances are much more extensive than currently suggested by the guideline and should enable mass balance calculations at the beginning and end of each test block and the derivation of ECx and NOEC values on the basis of geometric mean measured concentrations the relevant compartment. This means that exposure verification in all treatments at the beginning and end of each test block is required at the very minimum (for the most relevant compartment). The requirement for mass balance calculations should be explained more and also calculation advise be given or examples provided.**

Rational: Ambiguity/ unclarity on what is required, particularly for fast degrading substances, which may potentially invalidate the test.



Agreement justifications:

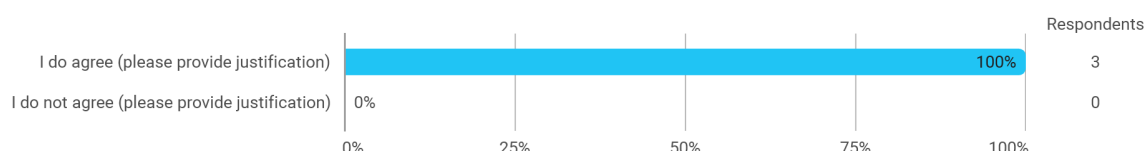
- However, historical studies should be still taken into consideration

Disagreement justifications

- the characteristics of a test item should always be taken into consideration with regard to testing scenario, sampling schedule and evaluation. No need to cover all possibilities in the TG.
- Disagree: It should not be expected that all types of substances can be stable over such a long test in static conditions. This is not realistic. In addition, geometric mean concentrations should not be requested, they should even be avoided, because it is impossible reflect the real exposure of the test organisms via complex routes including food. The best way to express the results is to use the nominal concentrations. In addition, the performance of a Life Cycle test is only reasonable for relatively stable compounds. Otherwise, an OECD 218 or 219 should be sufficient. The test is not frequently performed and will only be used in rare

cases. This should not be the case for rather instable compounds for which no chronic exposure in the environment has to be expected.

3. **para 38 - Despite the suggestion to change the text further to avoid ambiguity on what might be needed (see above comment), the following text elements should anyway be adapted to 'As a minimum, samples of the overlying water, pore water and the sediment should be analysed at the start of exposure (in case of water spiking preferably one hour after application) and at the end of each testing block, at the highest and lowest test concentration. This applies to vessels from both generations.**
4. **Rational: Using at the end of the experiment (instead of each testing block) suggests that no dose verification must be performed in between, which is not the case. Suggesting a 'lower concentration' is risky, because, if effects are seen in concentrations lower than the one chosen, these results cannot be used.**

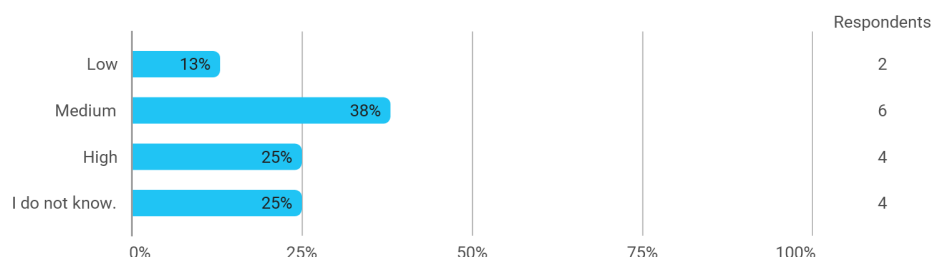


Agreement justifications:

- would be clearer
- agree

C.19 Test No. 235: Chironomus sp., Acute Immobilisation Test

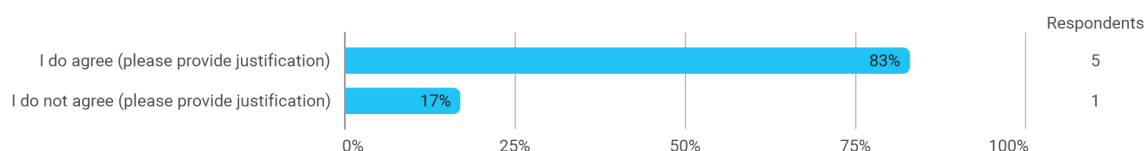
Relevance of revision



Interest in being part of an OECD Expert Group



1. **para 17 - it has been experienced that it is mandatory to transfer small amounts of fish food to the test vessels, it would even be better to state a feeding of the organisms on the day of application.**



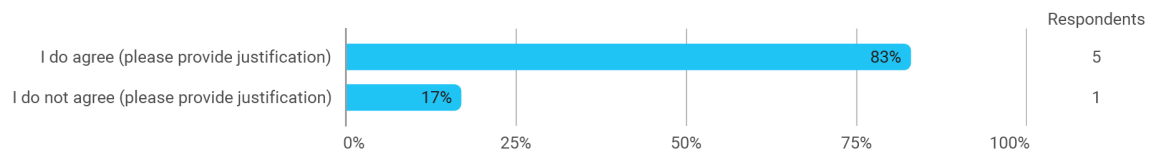
Agreement justifications:

- ▶ survival is low after 48h without feeding
- ▶ Control mortality can be too high without any food. When inserting the larvae to the test, very small amount of food are transferred from preculture to the test. Would be better to make this point clearer or to feed the test itself.
- ▶ agree
- ▶ No feeding is necessary in our experience

Disagreement justifications

- ▶ Feeding is not necessary within the 48 h testing period.

2. para 19 - The test may be carried out using semi-static renewal or flow-through system when the concentration of the test substance is not stable.



Agreement justifications:

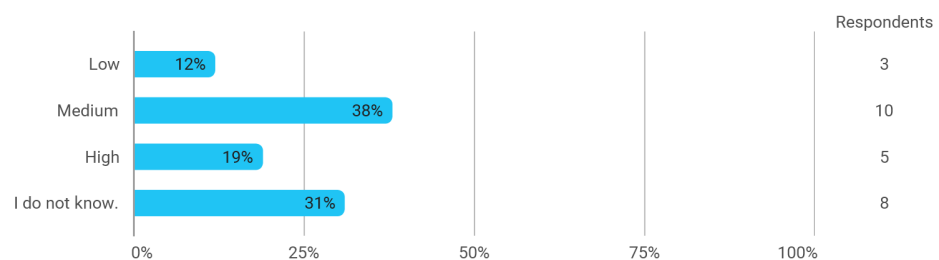
- ▶ could be mentioned
- ▶ should be mentioned
- ▶ I think a flow-through may cause additional complexity and costs for little value, but agree with using a semi-static renewal.
- ▶ Yes, for certain substances required

Disagreement justifications

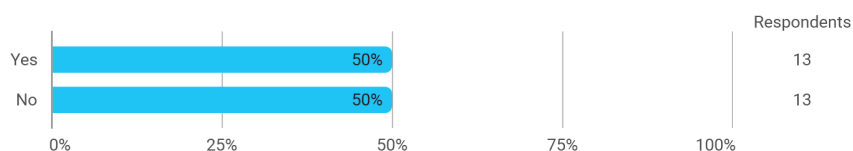
- ▶ Not necessary to be mentioned in the guideline. Refer to OECD GD 23.

C.20 Test No. 221: Lemna sp. Growth Inhibition Test

Relevance of revision



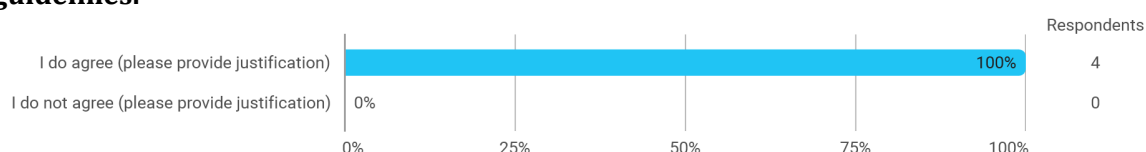
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. Include a section on initial considerations in line with the latest update of OECD TG 203 (18 June 2019).

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.

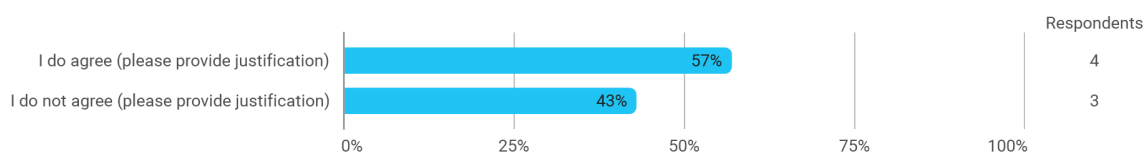


Agreement justifications:

- ▶ Align with other TGs
- ▶ basically agree, but the initial considerations should be well adapted to macrophyte testing. Also, several aspects mentioned in the initial considerations section of OECD 203 (2019) are clear as it is in the interest of the study performer to ensure exposure of the test organism to the compound and for this, considering important phys.-chem. data. No urgency.
- ▶ agree

2. Add the following text from OECD TG 203: When testing difficult test chemicals, Guidance Document No. 23 (OECD, 2019) should be consulted and the study design modified appropriately. For volatile and other difficult test chemicals, further specific measures should be taken (OECD, 2018). Any silicone materials that were in contact with the test solution(s) should be preferably discarded and not reused for subsequent tests with different test chemicals.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

- ▶ to exclude possible contaminations of further tests

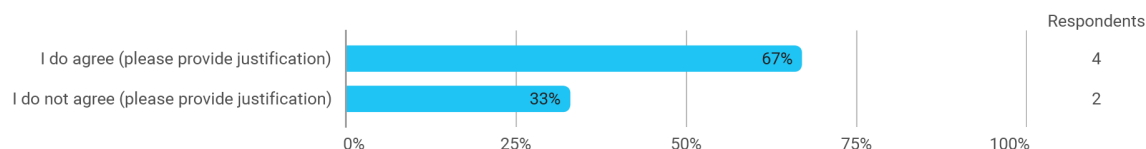
- Align with other TGs
- agree

Disagreement justifications

- This is addressed in section 11. Silicone materials usually are not necessary to be used in Lemna tests (contrary to fish tests).
- In general scientist should refer to the OECD GD 23 when testing difficult substances.

3. Test medium: clarify whether the use of natural water is applicable and if yes states that TOC/DOC should be measured. It is proposed to use the following wording from OECD TG 203: If natural water is used, DOC or TOC and nitrate-content (NO₃) should be measured once prior to the test.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

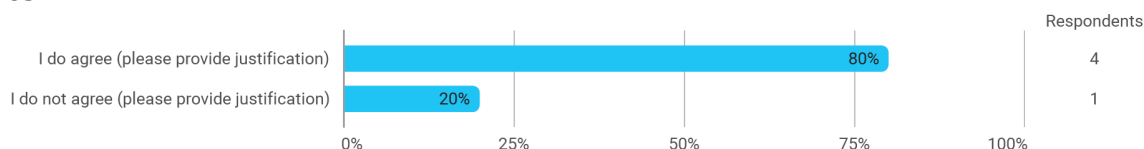
- to know the exact amount of organic carbon
- Align with other TGs
- agree

Disagreement justifications

- Annex 4 states that deionised water should be used; therefore, additional measurements are not required
 - Disagree: The growth media described in Annex 4 of this guideline are all prepared with deionised or distilled water. It is unclear why natural water should be added to the guideline.
4. para. 26: to be replaced by Para. 17 from OECD TG 203. The following text from Para. 23 of OECD TG 203 should also be included: Low toxicity solvents only (i.e. acetone, ethanol, methanol, tertiary-butyl alcohol, acetonitrile, dimethyl formamide, dimethyl sulfoxide, and triethylene glycol) as recommended in Guidance Document No. 23 (OECD, 2019) should be used whilst solvents of unknown toxicity should not be used. It should be noted that in spite of the low toxicity for fish, dimethyl formamide and dimethyl sulfoxide should be avoided where possible on human health and safety grounds.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium

characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

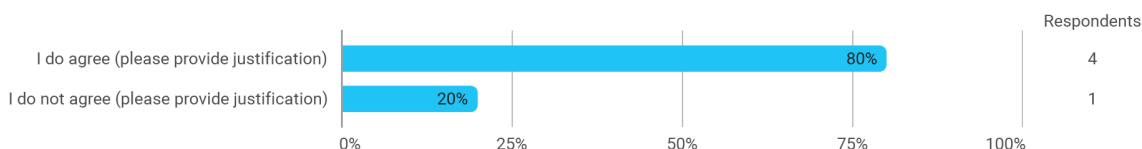
- ▶ to ensure the observed toxicity is not caused by the solvent
- ▶ Align with other TGs
- ▶ agree

Disagreement justifications

- ▶ DMF is a routinely used solvent in aquatic testing. It has been shown to be more suitable in most cases as e.g. acetone or ethanol. In general, untested substances, as routinely tested in ecotox labs, have to be handled as if they were CMR compounds. Thus at least the safety measures in the labs should always be appropriate. Based on the fact that DMF was predominantly used by CRO's most of the historical data being available for possible effects of solvents, relate to DMF. The currently ongoing project investigating whether always a solvent control and a control is needed when a solvent is used in a study based their evaluations as well on a high number of studies which have been performed with DMF. The hope is, that this evaluation will allow to skip the negative control and by that reduce animal numbers in vertebrate tests. If we lose DMF that might have a negative impact on this decision.

5. Add Para. 18 from OECD TG 203.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

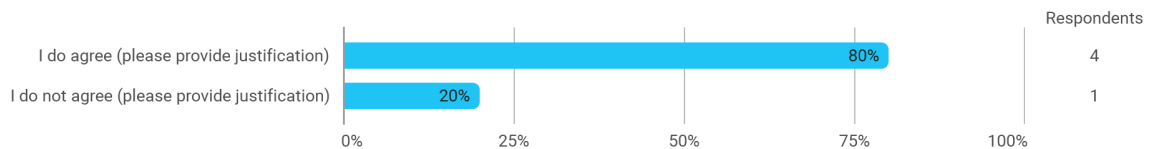
- ▶ makes sense
- ▶ Align with other TGs
- ▶ agree

Disagreement justifications

- ▶ Disagree: Paragraph 18 from OECD TG 203 cannot be simply copied to OECD TG 221 without adaptations e.g. of the given pH values and ranges. Should not limit the pH range from 6.0 to 8.5 but to the tolerances of the plants

6. Frequency of analytical determinations and measurements (Para. 42-47): align with Para 24-25 of OECD TG 203.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

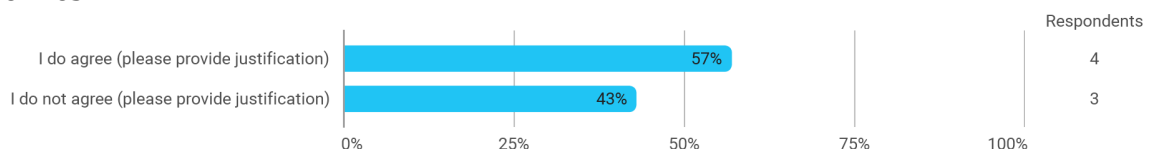
- ▶ only parameters that are needed for the test system
- ▶ Align with other TGs
- ▶ agree

Disagreement justifications

- ▶ Disagree: The paragraphs 42-47 of OECD TG 227 are clear and do not require modification. Particularly an alignment with paragraph 25 of OECD TG 203 is seen very critical: 'During the exposure, dissolved oxygen, pH, salinity (if relevant) and temperature should be measured daily in each test vessel'. The additional value of these daily measurements is not clear. They can even become detrimental as putting electrodes daily in almost sterile vessels massively increases the risk of contamination.

7. Treatment of results: Add the following text from Para. 31 of OECD TG 203: It is recommended that results should be calculated using the measured concentrations of the test chemical. If the deviation from the nominal concentrations is smaller than 20%, results may also be based on the nominal concentrations. It should be noted that it is often useful to have both measured and nominal effect concentrations quoted, see Guidance Document No. 23 (OECD, 2019).

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

- ▶ because it makes sense
- ▶ Align with other TGs

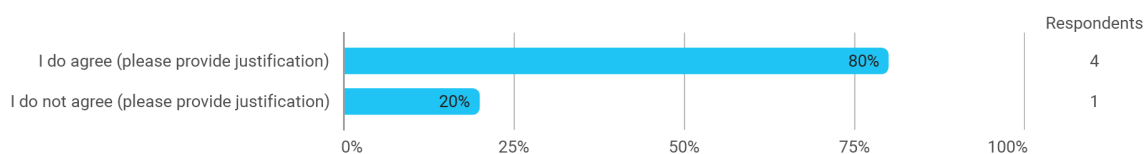
- agree

Disagreement justifications

- See para 47.
- Disagree: The procedures as described in paragraphs 45-47 of OECD TG 221 are clear and do not require any addition. With regard to the point that results may be presented on nominal and on measured concentrations in parallel, there is the risk that two complete statistical evaluations and two sets of endpoints finally produce more confusion than value. No need for a change.

8. Include the Annex 3 from OECD TG 203 or at least include consistent information related to the content in particulate matter, TOC and COD in the description what makes an acceptable dilution/test water.

Rational: The text from OECD TG 203 provides a more explicit reference to regulatory requirements related to difficult to test substances as well as requirements related to the validation of analytical methods and the validity criteria related to test medium characteristics. This text should be consistent across relevant aquatic toxicity test guidelines.



Agreement justifications:

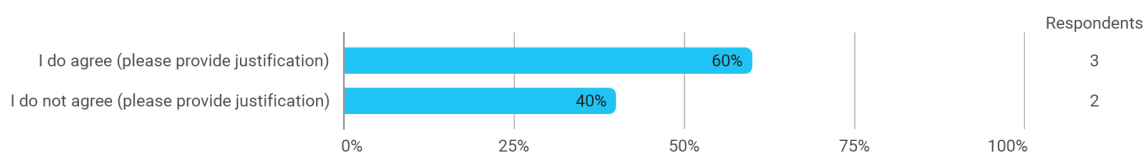
- to have an orientation what is tolerable
- Align with other TGs
- agree

Disagreement justifications

- Disagree: The growth media described in Annex 4 of this guideline are all prepared with deionised or distilled water, thus water of no or extremely low contents of particulate matter and TOC. Giving ranges for / measuring these variables is therefore not considered necessary.

9. para 34 - The only way of exposure in the test is by mixing the substance in the water. However, especially for contact herbicides, there are strong indications that this is not the worst-case route of exposure. Direct contact of the floating plant parts to the substance might cause higher toxicity. Hence, an overspray scenario should be implemented in the test guideline. It is understood that currently reference is made to publication 11 by Lockhart 1989, but it is being wondered if the guideline can be updated with such a very relevant scenario.

Rational: Lack of guidance, referred publication is old



Agreement justifications:

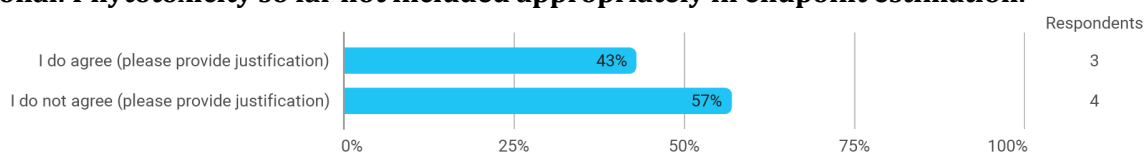
- in order to get a better impression on the impact on the plants under realistic conditions
- agree

Disagreement justifications

- Endpoints are intended for use in risk assessments in which ECx values are compared with predicted environmental concentrations. An overspray scenario in a Lemna test would result in exposure via foliage and water. Deriving a meaningful numerical endpoint from a combined exposure route is difficult and cannot readily be compared with a corresponding PEC value. If contact activity is a concern, then compounds can be tested in terrestrial plant studies, where foliar application is standard
- Disagree: The guideline clearly states that this application method is not covered, adding it would mean that a ring test is needed. The method has not been validated and standardized for this methodology. Before this application method could be included in a guideline the necessary steps as described for new guidelines under OECD have to be considered.

10. paras 38, 65 – The OECD 221 guidance requires the observation and reporting of (signs of) phytotoxicity, but guidance on how to do that and also on how to interpret these observations and consider them in the endpoint derivation is lacking. Phytotoxic effects (particularly if a dose response pattern of phytotoxicity can be observed) should be covered by the relevant endpoint for risk assessment, i.e, phytotoxicity should be at the maximum be present for 50% of the test fronds/biomass at the ErC50 (or EyC50 as this separate issue is not yet solved either, both endpoints should be reported). Phytotoxicity should be observed and differentiated at the minimum in terms of the exact observation (necrosis, chlorosis, short roots, etc.... see paragraph 38), each on an ordinal scale (+, ++, +++), but preferentially on a percental scale, so that incidence and severity are quantified. This allows to recommend an endpoint for risk assessment that clearly covers phytotoxicity sufficiently. A section on this could be added to the chapter 'Response variables', which currently does not address the inclusion of phytotoxicity observations.

Rational: Phytotoxicity so far not included appropriately in endpoint estimation.



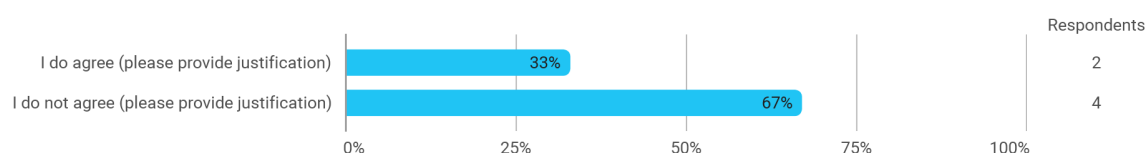
Agreement justifications:

- it is already expected from the authorities, but still missing in guidelines to be mentioned
- agree

Disagreement justifications

- ▶ Changes in plant development are only supporting information and for a good reason: these parameters are much more subjective observations than frond numbers, especially when it comes to judging how many percent are affected and thus it is questionable if this data would be reproducible on an intra- and inter-laboratory scale.
- ▶ Visual effects on Lemna invariably lead to reductions in biomass and frond number. These parameters are already assessed and provide a qualitative and objective assessment of effects. It is uncertain whether estimation of endpoints based on visual observations will add value to risk assessment and this point is currently being clarified by a WG working under SETAC Plant Interest Group.
- ▶ Disagree: A semi-quantitative assessment of phytotoxicity with an ordinal scale may be added to ensure that severe effects are not overlooked when deriving numerical endpoints for frond number and (e.g.) dry weight. However, quantifying phytotoxicity on a percental scale would require alignment and confirmation in a ring test before being implemented (aligned list of symptoms and reproducible quantitative estimates). For instance, what would 50% necroses mean? A) 50% of all fronds with (at least some) brown spots; b) 50% of all fronds brown/dead; c) fronds damaged by 50% of their surface (otherwise green and healthy)... In practice, the estimation of damage on a percental scale proved to be rather difficult and of limited reproducibility.
- ▶ Changes in plant development are only supporting information and for a good reason: these parameters are much more subjective observations than frond numbers, especially when it comes to judging how many percent are affected and thus it is questionable if this data would be reproducible on an intra- and inter-laboratory scale.

11. para 48 - limit test in the case of tests with formulated products, it would be appropriate to specify whether the concentration refers to the product or to the active ingredient (or one of the active ingredients).



Agreement justifications:

- ▶ obviously needed to be addressed in the report if based on test item (formulation) or active ingredient

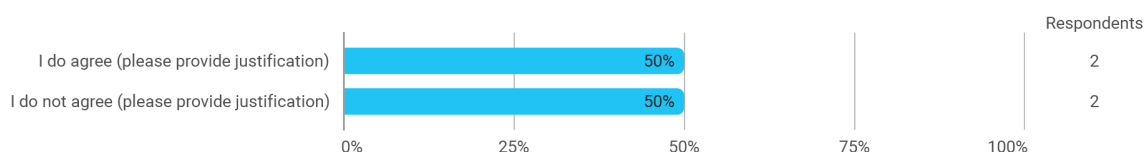
Disagreement justifications

- ▶ Generally, it is 100 mg test item/L and accordingly if the test item is a formulation, it is 100 mg formulation/L. Details may depend on the purpose of the study but this document is not the right place to give guidance in this regard (not a Lemna test system specific question).
- ▶ This is not needed since this is dependent on the way the endpoints are used and the regulatory framework.
- ▶ this can cause confusion in case of mixtures or formulations with multiple a.i.s with different ratios. This should be included in a separate guideline on mixtures

- Generally, it is 100 mg test item/L and accordingly if the test item is a formulation, it is 100 mg formulation/L. Details may depend on the purpose of the study but this document is not the right place to give guidance in this regard (not a Lemna test system specific question).

12. para 41 - “Fresh weight”: Suggested update: Alternatively or in place of the sentence “All colonies are transferred to pre-weighed polystyrene tubes (...)”, include the indication “All colonies are collected from each of the test vessels and rinsed with distilled or deionised water. They are blotted to remove excess water and the fresh weight is determined by weighing.”

Rational: This procedure is the same of the first part of preparation for the dry weight determination in this guideline and the procedure indicates for *Myriophyllum* test (OECD 238, par. 41). Drying on absorbent paper is sufficient to remove the water, while the transfer of the material into the centrifuge tubes makes recovery difficult if the determination of dry weight is also required (especially when the plants show necrosis and chlorosis).



Agreement justifications:

- makes sense

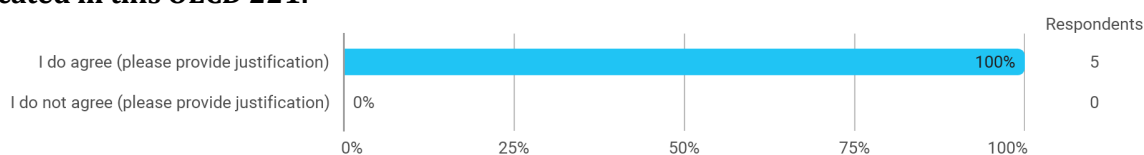
Disagreement justifications

- Disagree, no need for a change.
- blot drying might not be sufficient to remove the amount of water/medium that can affect measurements

13. Annex 4: Pag. 20, 20xAAP growth medium: update of the table changing the concentration in stock solution and concentrations in prepared medium based on AAP composition (OECD 201 and OCSPP850.4400). Moreover, referring to the note below the table in the guideline, the possibility of using a medium with NaHCO₃ concentration at 15 mg/L could be added; the medium would be easier to prepare for dissolving the salts.

Rational: The concentrations of the stocks and the corresponding concentrations in the medium differ (often for reasons of approximation) from those of the medium AAP reported in OECD 201 and of the medium 20xAAP reported in OCSPP 850.4400.

Moreover, for medium AAP the salt K₂HPO₃ is required, instead of K₂HPO₃·3H₂O indicated in this OECD 221.



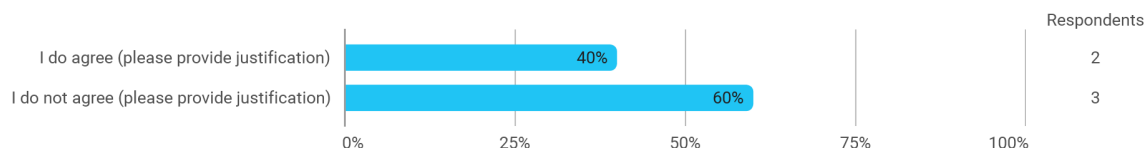
Agreement justifications:

- makes sense

- Agreed on the first part, consistency in between the guidelines would be desirable. Doubtful if the second part (Moreover,...) is appropriate without further ringtesting and if the benefit of easier preparation would justify this modification.

- agree

14. The methodology as described in the current guideline is not suitable to test for effects of virtually insoluble particulate materials (e.g. nano-forms of substances, polymers). Adaption of the guidelines methodology with latest scientific findings is indicated so that valid tests with particulate materials can be performed.



Agreement justifications:

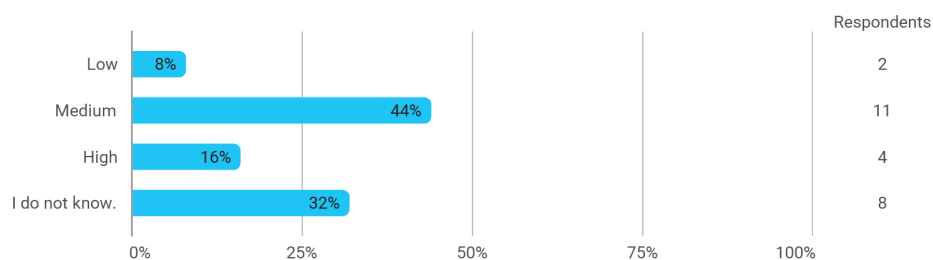
- makes sense

Disagreement justifications

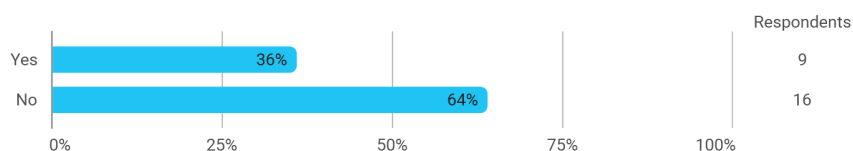
- Not Lemna-specific, rather to be included in OECD 23.
- this should be described in a separate guideline as methodology is not sufficiently established yet

C.21 Test No. 238: Sediment-Free Myriophyllum Spicatum Toxicity Test

Relevance of revision



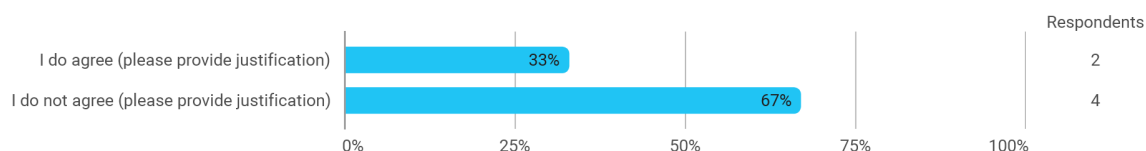
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. para 2 - Guidelines on other (rooted, submerged or semi-aquatic) macrophyte species is required , particularly on the interpretation of the validity criteria when species are tested that were not validated or considered in the OECD guideline. Is it possible to include other species under this guideline and make this a more general guideline, or are other species specific guidelines required?

Rational: Current validity criteria do not always allow for other species to provide a valid test under this guideline.

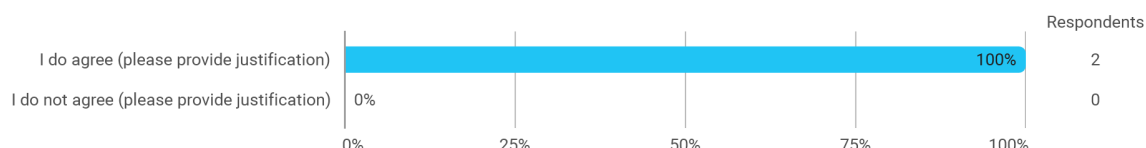


Disagreement justifications

- Other species are usually tested under TG 239 as some of them require sediment and it is very difficult to have other plants in sterile culture.
- depends on the species
- This TG is specific to *Myriophyllum spicatum* and due to test system constraints (test tubes &) cannot readily be adapted to for emergent macrophytes or rooted macrophytes with floating leaves. There is already an OECD TG under development for the emergent macrophyte, *Glyceria maxima*
- Disagree: Further species should not be included into a guideline without having proven their suitability and reliability in a ring test. This applies particularly to macrophytes with frequently huge variation in growth and different growth forms for the same species. For a potential extension of sediment-free test methods only species not rooting in the sediment should be considered to avoid production of artifacts. Before other species can be added their suitability has to be checked and the method needs respective standardization and validation.

2. **para 22 - Add a sentence specifying that “the final concentration of nutrients in the test solutions should be the same reported in the table in Annex 2 (with sucrose at 3%)”**

Rational: it is unclear how the solutions should be prepared and it could be interpreted that the medium to be used for the test solutions (see also par. 21) is with 30% sucrose. Instead, in paragraph 20 and also in ASTM it is specified that the medium for culturing and for testing is the same. Furthermore, it could be envisaged the possibility of preparing the solutions directly in the medium, as in other guidelines (in this case the tenfold medium would not be used).

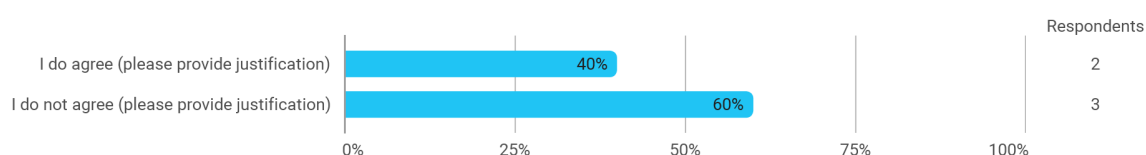


Agreement justifications:

- just for clarification
3. **paras 36, 63 - The OECD 239 guidance requires the observation and reporting of (signs of) phytotoxicity and root development, but guidance on how to do that and also on how to interpret these observations and consider them in the endpoint**

derivation is lacking. Phytotoxic effects (particularly if a dose response pattern of phytotoxicity can be observed) should be covered by the relevant endpoint for risk assessment, i.e, phytotoxicity should be at the maximum be present for 50% of the test fronds/biomass at the ErC50 (or EyC50 as this separate issue is not yet solved either, both endpoints should be reported). Phytotoxicity should be observed and differentiated at the minimum in terms of the exact observation (necrosis, chlorosis, short roots, etc.... see paragraph 63), each on an ordinal scale (+, ++, +++), but preferentially on a percental scale, so that incidence and severity are quantified. This allows to recommend an endpoint for risk assessment that clearly covers phytotoxicity sufficiently. A section on this could be added to the chapter 'Response variables', which currently does not address the inclusion of phytotoxicity observations.

Rational: Phytotoxicity so far not included appropriately in endpoint estimation.



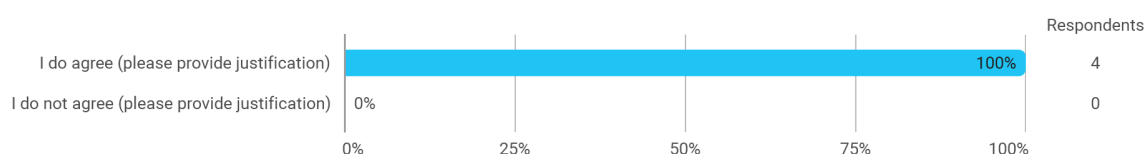
Agreement justifications:

- absolutely needed

Disagreement justifications

- it's just accompanying observations, no need to quantify them. Statistical evaluation wouldn't be feasible. Observations are subjective, not measurable values.
- Visual effects invariably translate into effects on biomass and shoot length. These parameters are already assessed and provide a qualitative, objective assessment of effects. In contrast, visual scores are subjective, lack granularity and do not provide a measure of variability. It is also uncertain whether estimation of endpoints based on visual observations will add value to risk assessment given that this TG already generates 6 endpoints. The issue of visual score endpoints is currently being clarified by a WG working under SETAC Plant Interest Group.
- Disagree: A semi-quantitative assessment of phytotoxicity with an ordinal scale may be added to ensure that severe effects are not overlooked when deriving numerical endpoints for weight and length. However, quantifying phytotoxicity on a percental scale would require alignment and confirmation in a ring test before being implemented (e.g. % of individuals affected? % of plant surface or plant length affected?). It should further be noted that not all symptoms of phytotoxicity can be expressed quantitatively (e.g. discoloration, curved shoots).

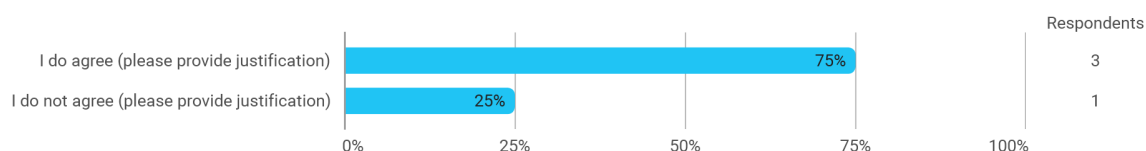
4. paras 37 and 41 - clarify if "fresh weight" means shoot fresh weight or total plant fresh weight; same for dry weight



Agreement justifications:

- definition not clear
- for better understanding
- agree

5. para 45 - Should it not read 80% instead of 20% in these paragraphs (the allowed variation is $\pm 20\%$, but the concentrations should not be 20% of nominal).



Agreement justifications:

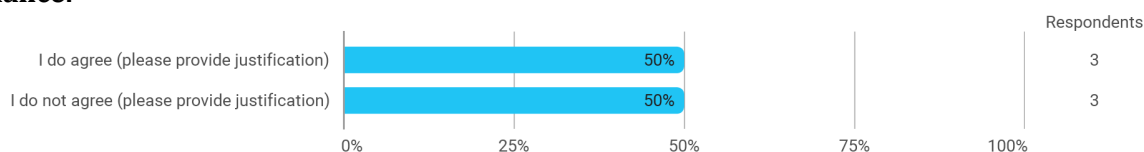
- Correction needed

Disagreement justifications

- current wording seems correct

6. para 57 - The paragraph addresses the controversy of the ErC50 vs the EyC50, which in itself is appreciated. However, it causes confusion with regard to certain publications, where the ErC50 has been compared to mesocosm results and found to not always be protective. Can the paragraph be revised in the light of these scientific findings?

Rational: This is for a long time a topic of discussion in the risk assessment world and would need to be addressed by both the study guidelines and the risk assessment guidance.



Agreement justifications:

- would make sense
- agree

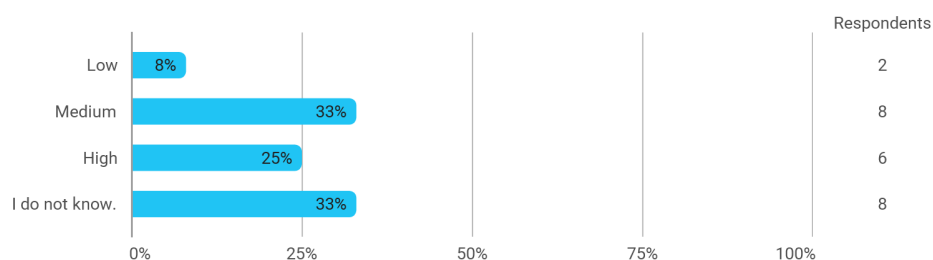
Disagreement justifications

- The use of an ErC50 v EyC50 is a risk assessment question and does not influence the performance of the test.
- The use of ErC50 v EyC50 endpoints is a risk assessment question and the preference varies between countries and risk assessment schemes. Therefore, the question of protectiveness is not appropriate within a global (OECD) TG.
- Disagree: The paragraph is correct and does not need any revision. According to the guideline both growth rate and yield endpoints shall be calculated and reported. The final endpoint selection for risk assessment is not a guideline topic. The comment with respect to

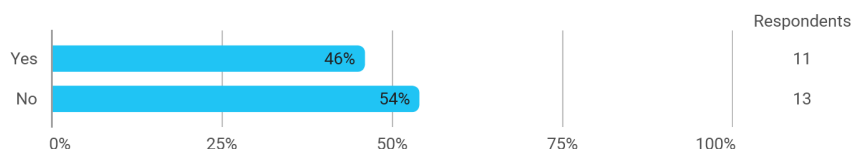
comparison with mesocosm studies is wrong! The mentioned publication to which this comment is referring to, is an oral presentation from UBA at the SETAC Rome. This was never published in a peer reviewed journal. Publications in peer reviewed journals as. e.g. by Arts & van Wyngarden came to a different result. The underlying data from the UBA evaluation were never made publicly available. E.g. the aquatic ad hoc group under the umbrella of CLE asked for the data but only received the presentation slides which do not contain the necessary information. The statement that the tier 1 risk assessment is not protective therefore is in contradiction to real publication and other presentations at SETAC where it was presented that the tier 1 risk assessment for primary producers is protective in most cases. Therefore, this comment is wrong and misleading! Anyhow this is a statement which should not be part of a guideline at all. A Guideline describes a test methodology. The regulatory framework is not part of a guideline.

C.22 Test No. 239: Water-Sediment Myriophyllum Spicatum Toxicity Test

Relevance of revision



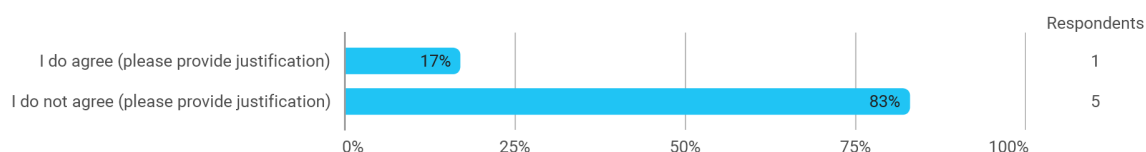
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

- para 2 - Guidelines on other (rooted, submerged or semi-aquatic) macrophyte species is required , particularly on the interpretation of the validity criteria when species are tested that were not validated or considered in the OECD guideline. Is it possible to include other species under this guideline and make this a more general guideline, or are other species-specific guidelines required?**

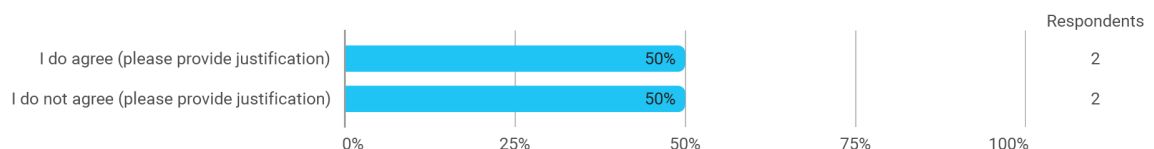
Rational: Current validity criteria do not always allow for other species to provide a valid test under this guideline.



Disagreement justifications

- ▶ other species may have slightly different requirements. A sentence could be added that testing of other species is possible but some parameters might have to be adapted, depending on the requirements of the alternative plant species.

- ▶ Other macrophytes as glyceria are mentioned. It would be a very long list to mention all. Some other macrophytes maybe need longer time for doubling.
 - ▶ depends on the tested species
 - ▶ This TG is specific to *Myriophyllum spicatum* and can be adapted for other completely submerged macrophytes with similar morphology. It cannot be readily adapted for emergent macrophytes or rooted macrophytes with floating leaves. There is already an OECD TG under development for the emergent macrophyte, *Glyceria maxima*
 - ▶ Disagree: Further species should not be included into a guideline without having proven their suitability and reliability in a ring test. This applies particularly to macrophytes with frequently huge variation in growth and different growth forms for the same species. Considering that with *Glyceria maxima* an OECD guideline on a third macrophyte species is expected soon, there is currently no need for additional species-specific guidelines.
2. **paras 4, 77 - Should it not read 80% instead of 20% in these paragraphs (the allowed variation is $\pm 20\%$, but the concentrations should not be 20% of nominal)?: 'In semi-static tests (i.e. exposure via the water phase) where the concentration of the relevant test chemical(s) is not expected to remain within 80% of the nominal concentration over the test duration without renewal of test solutions, used and freshly prepared test solutions should be sampled for analyses of test chemical concentration at each renewal.'**



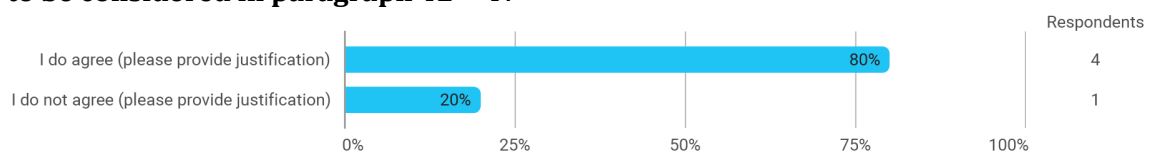
Agreement justifications:

- ▶ Correction needed

Disagreement justifications

- ▶ Paragraph seems right

3. **paras 30 and 32 - Include a "Test Design C: one shoot per test vessel" in paragraph 32; with 10 replicates for control(s) and 5 replicates for concentration levels in paragraph 30; also to be considered in paragraph 42 + 47**



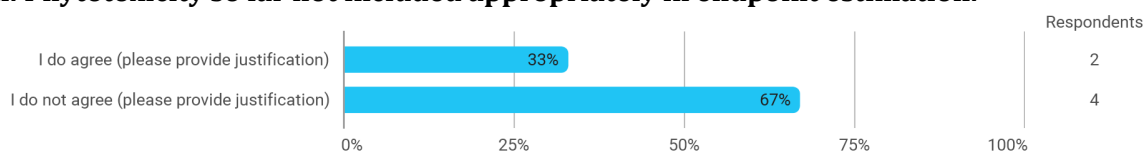
Agreement justifications:

- ▶ In line with EFSA supporting publication 2019:EN-1673
- ▶ alternative test designs should be mentioned in the other passages
- ▶ for clarification and because this design works without pseudoreplicates!

Disagreement justifications

- ▶ This test design is covered under para 32 which states: “Alternative test designs of one shoot per pot per test vessel are acceptable provided that replication is adjusted as required to achieve the required validity criteria.”
- 4. **paras 63, 94 - The OECD 239 guidance requires the observation and reporting of (signs of) phytotoxicity and root development, but guidance on how to do that and also on how to interpret these observations and consider them in the endpoint derivation is lacking. Phytotoxic effects (particularly if a dose response pattern of phytotoxicity can be observed) should be covered by the relevant endpoint for risk assessment, i.e, phytotoxicity should be at the maximum be present for 50% of the test fronds/biomass at the ErC50 (or EyC50 as this separate issue is not yet solved either, both endpoints should be reported). Phytotoxicity should be observed and differentiated at the minimum in terms of the exact observation (necrosis, chlorosis, short roots, etc.... see paragraph 63), each on an ordinal scale (+, ++, +++), but preferentially on a percental scale, so that incidence and severity are quantified. This allows to recommend an endpoint for risk assessment that clearly covers phytotoxicity sufficiently. A section on this could be added to the chapter ‘Response variables’, which currently does not address the inclusion of phytotoxicity observations.**

Rational: Phytotoxicity so far not included appropriately in endpoint estimation.



Agreement justifications:

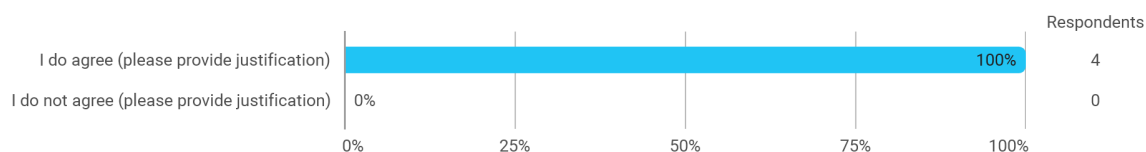
- ▶ absolutely needed

Disagreement justifications

- ▶ it's just accompanying observations, no need to quantify them. Statistical evaluation wouldn't be feasible. Observations are subjective, not measurable values.
- ▶ Observations are always a bit subjective. This can not be statistically evaluated. To get hard and reliable data, it would be necessary to measure, for example, root length or chlorophyll this makes only sense in very special cases.
- ▶ Visual effects invariably translate into effects on biomass and shoot length. These parameters are already assessed and provide a qualitative, objective assessment of effects. In contrast, visual scores are subjective, lack granularity and do not provide a measure of variability. It is also uncertain whether estimation of endpoints based on visual observations will add value to risk assessment given that this TG already generates 6 endpoints. The issue of visual score endpoints is currently being clarified by a WG working under SETAC Plant Interest Group.
- ▶ Disagree: A semi-quantitative assessment of phytotoxicity with an ordinal scale may be added to ensure that severe effects are not overlooked when deriving numerical endpoints for weight and length. However, quantifying phytotoxicity on a percental scale would require alignment and confirmation in a ring test before being implemented (e.g. % of individuals

affected? % of plant surface or plant length affected?). It should further be noted that not all symptoms of phytotoxicity can be expressed quantitatively (e.g. discoloration, curved shoots).

5. para 69 - change text to "or within one hour after application for substances that are not stable"

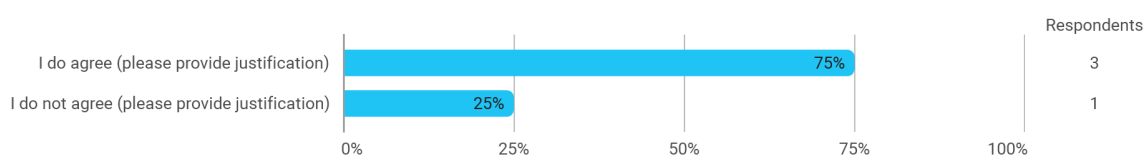


Agreement justifications:

- ▶ current wording doesn't match for fast degrading substances
- ▶ Yes, because for unstable substances it makes no sense to wait for 1 hour
- ▶ makes sense

6. paras 71, 73, 75, 79 - The paragraphs on analytical verification seem to create confusion on measurements are required, particularly when formulations with (multiple) active substance that are (all) more or less fast degrading are tested. It needs to be clarified that paragraph 79 is the leading principle, leading to the need to provide measurements for all treatments in order to derive endpoints based on geometric mean measured concentrations and that also for these tests a mass balance is required. The consequences for the required analytical verifications need to be made clearer. Please also refer to the comments provided on revisions for OCED 238, 233, OECD 218 and 219.

Rational: Ambiguity/ unclarity on what is required, particularly for fast degrading substances, which may potentially invalidate the test.



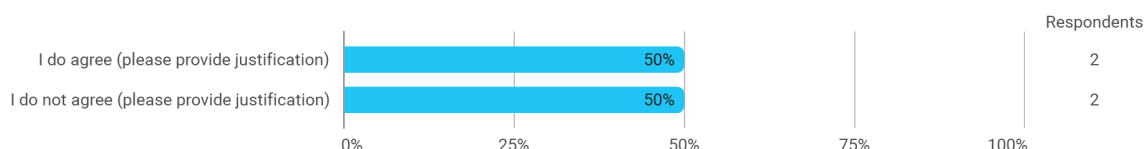
Agreement justifications:

- ▶ would be very helpful
- ▶ Partly agreed: The rules for analyzing all treatment groups or only the highest and the lowest treatment level may be described clearer.

Disagreement justifications

- ▶ In usual spiked water applications, only traces of test item can be found in sediment and pore water. Sampling and analysing sediment and pore water is completely unnecessary and won't make much difference in the mass balance. It could be noted though in the TG that this approach might be followed in case of applications via spiked sediment or in case the substance is known to strongly absorb to sediment.

7. **para 89- The paragraph addresses the controversy of the ErC50 vs the EyC50, which in itself is appreciated. However, it causes confusion with regard to certain publications, where the ErC50 has been compared to mesocosm results and found to not always be protective. Can the paragraph be revised in the light of these scientific findings? Rational: This is for a long time a topic of discussion in the risk assessment world and would need to be addressed by both the study guidelines and the risk assessment guidance.**



Agreement justifications:

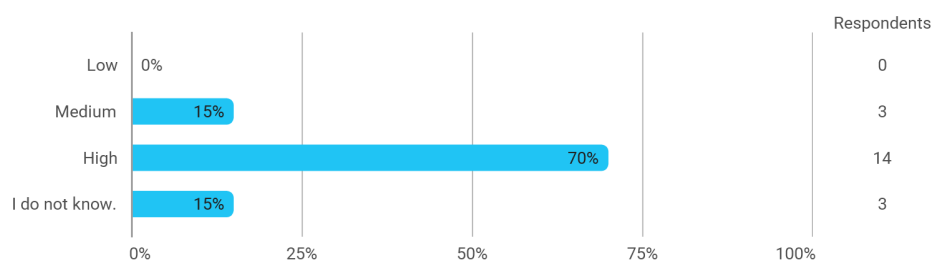
- would be very helpful

Disagreement justifications

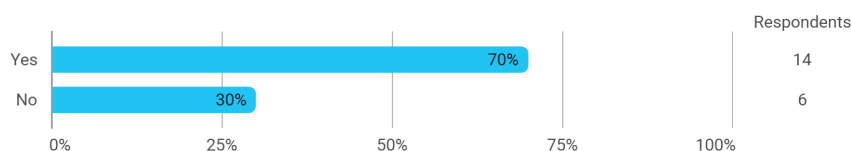
- the use of ErC50 v EyC50 endpoints is a risk assessment question and the preference for EyC50 v ErC50 endpoints varies between countries and risk assessment schemes. Therefore, the question of protectiveness is not appropriate within the context of a global (OECD) TG.
- Disagree: The paragraph is correct and does not need any revision. According to the guideline both growth rate and yield endpoints shall be calculated and reported. The final endpoint selection for risk assessment is not a guideline topic. The comment with respect to comparison with mesocosm studies is wrong! The mentioned publication to which this comment is referring to, is an oral presentation from UBA at the SETAC Rome. This was never published in a peer reviewed journal. Publications in peer reviewed journals as e.g. by Arts & van Wyngarden came to a different result. The underlying data from the UBA evaluation were never made publicly available. E.g. the aquatic ad hoc group under the umbrella of CLE asked for the data but only received the presentation slides which do not contain the necessary information. The statement that the tier 1 risk assessment is not protective therefore is in contradiction to real publication and other presentations at SETAC where it was presented that the tier 1 risk assessment for primary producers is protective in most cases. Therefore, this comment is wrong and misleading! Anyhow this is a statement which should not be part of a guideline at all. A Guideline describes a test methodology. The regulatory framework is not part of a guideline.

C.23 Test No. 231: Amphibian Metamorphosis Assay

Relevance of revision



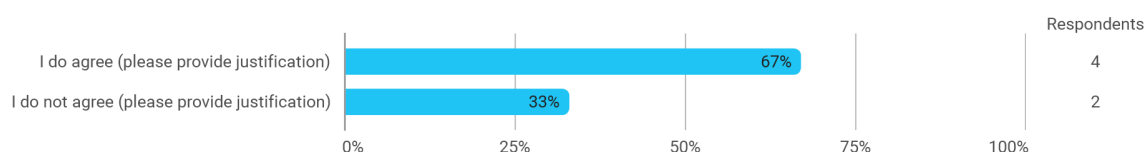
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

- para 9 - Suggested addition highlighted in bold: Measured aqueous iodide concentrations from the test water should be reported. Iodide should be measured from test vessels from one replicate of the control and the highest test concentration at the beginning and end of the test to ensure sufficient iodide, and to ensure that the test item is not interfering with iodide concentrations.....Based on the available data from the validation studies, the protocol has demonstrated to work well when test water iodide concentrations ranged between 0.5 and 10 ug/L.**

Rational: Current TG states that "measured aqueous iodide concentrations from the test water should be reported" and includes the recommended range (0.5-10 ug/L). However there is no further guidance on how, how frequently or where these measurements should be taken from. This is being interpreted differently by different CROs and Industry partners. Some CROs are adding a given amount of iodide and since its stable, are not seeing the need to measure it. Some CROs are measuring iodide every 6 months from central water source. And some are measuring more routinely in study vessels. Given the importance of sufficient iodide in order for the thyroid gland to synthesise TH, and excessive iodide concentrations could potentially mask the effects of thyroid active substances (stated in OPPTS 890.1100) more clarity and further guidance is needed, as suggested above.



Agreement justifications:

- it should demonstrate that iodide is also available in exposed replicates
- Update needed

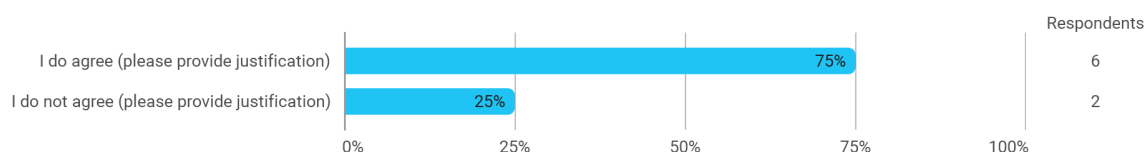
Disagreement justifications

- In the case of artificial test water, not necessary because defined quantities are added
- If iodide concentrations are measured in the diet, then it should not be required to measure concentrations in the test water. In addition, characterization of the source water (and diet) on a yearly basis should be sufficient. Testing the high test level is not warranted.

- para 13 - Current feeding rations recommended in Table 1 should be reduced to around 50%.**

Rational: The current TG recommendation for feeding regime is too high and leads to >20% late stage tadpoles (> NF60) which require special complex statistical analysis and can confound the interpretation of data. Reducing the recommended feeding to

50% of current TG requirements would prevent fast development of tadpoles, avoid > 20% late stage individuals and help to reduce bacterial growth, meaning test item stability and DO is easier to maintain. This would also lead to reduced need for cleaning and less stress to tadpoles.



Agreement justifications:

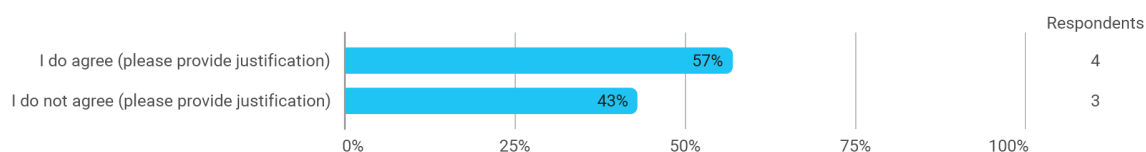
- ▶ These feeding rates seem to be satisfying
- ▶ There is good practical experience with reduced feeding, less pollution, less >NF60
- ▶ update needed
- ▶ this common practise so should be reflected to ease regulatory acceptance and deliver better studies

Disagreement justifications

- ▶ Many labs would likely experience limited growth during the study leading to failure to identify any effects
- ▶ Laboratories conducting the test should be able to determine what works best in their particular situation. For example, flow rates can vary in different labs and among different studies resulting in variability in feed being cleared from tanks. Thus, not being overly prescriptive here would be helpful. The key aspect is that tadpole growth and development meets validity criteria in the control group.

3. **para 17 - Suggested amendment below in bold: "If species specific acute toxicity data are not available, then a 14 day range finding test can be completed with tadpoles that are respective of those in the AMA (i.e. initiated with NF 51 tadpoles)....A minimum of 10 tadpoles should be used per treatment group"** Recommendations on the use of controls in range finders would also be beneficial i.e. use of a solvent control only, omitting the use of a clean water control in the interest of reducing animal use, should be acceptable.

Rational: 96 hour range finding studies are too short for MTC approximation. During longer (i.e. 21 day) exposures, tadpoles experiencing initial toxicity during the first few days of exposure can demonstrate recovery later on, which can confound the interpretation of the MTC. Alternatively, the opposite may be true for some compounds, whereby systemic toxicity effects are seen later on in chronic studies. A 14-day range finder is typically required as a minimum (potentially 21 days for some compounds). There is currently no guidance on recommended number of animals for range finding studies. Clearer guidance will help CROs justify animal use and obtain sufficient permits for animal use. Clearer guidance on the use of control groups during range finders is needed - i.e. it should be acceptable that a clean water control can be omitted and only a solvent control group used, in the interest of animal use.



Agreement justifications:

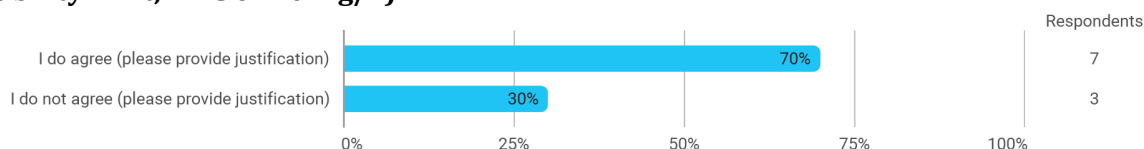
- Clean water medium not needed as solvents available have no (or low) toxicity. A 14 day or 7day (at least) range finder may avoid unexpected toxicity during the test - but I have no empirical data on this
- Agreed with a more detailed description of test design, however a 7 day exposure should be sufficient for acute data. If also chronic data are requested a prolonged exposure including a check for body weight or SVL might be meaningful

Disagreement justifications

- In general, I agree with an extended RF, however the duration should be up to expert judgement
- Do not agree – while the suggestions for conducting a rangefinder seem very appropriate, it is preferred not to be too prescriptive in directives for the rangefinder. Flexibility may be a better approach accommodating for different test materials with different amounts of existing data available.
- Enough flexibility should be given to allow reducing the use of animals

4. **para 16 - Suggested changes highlighted in bold below: "For the purpose of this test, the high test concentration should be set by the solubility limit of the test substance; the maximum tolerated concentration (MTC) for acutely toxic compounds; or 10 mg/L, whichever is lowest"**

Rational: Current TG sets the highest test concentration as the MTC, solubility limit or 100 mg/L, whichever is lowest. 100 mg/L is considered too high. The respective level 3 fish assays (OECD 229 and OECD 230) set a top concentration of 10 mg/L. There is a lack of convincing evidence to demonstrate that amphibians are 10 times less sensitive than fish, therefore 100 mg/L is considered too high. Recommendation to bring test concentrations back in line with level 3 testing for EAS modalities in fish (solubility limit, MTC or 10 mg/L).



Agreement justifications:

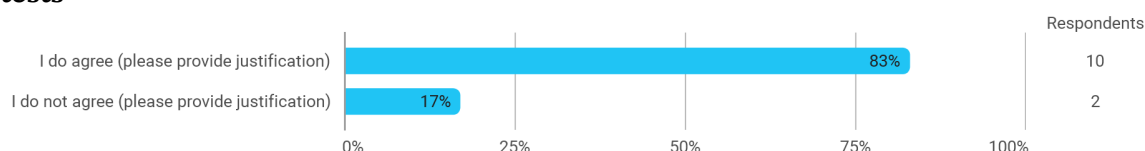
- 100 mg/L is a clear typo for a chronic study
- consistency with other TG
- This aligns with the fish short term reproduction assay.
- Alignment needed with other TGs (e.g., FSTRA, XETA)
- See comments on other TGs

Disagreement justifications

- ▶ I do not see why the concentration should be lowered if the test substance is still soluble and not systemically toxic at 100mg/L
- ▶ The current text in the guidance reflect the screening nature of the TG
- ▶ This should be discussed and agreed - if I remember well there is a rationale behind the choice of this limit concentration

5. paras 16, 17, 18 - Language around the MTC should be better defined and aligned across relevant aquatic endocrine TGs. The limit test level of 100 mg/L should be considered for consistency opposite similar TGs (limit levels of 10 mg/L).

Rational: To avoid regulatory confusion and potential for unnecessary repeat or higher tier tests



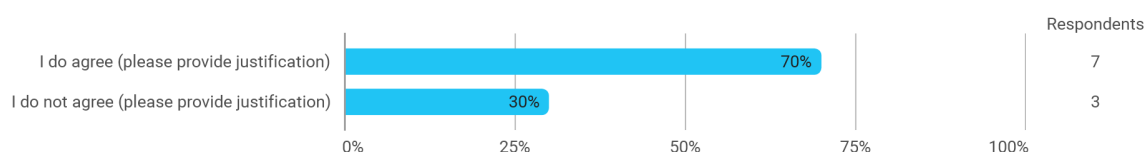
Agreement justifications:

- ▶ 100 mg/L is a clear typo for a chronic study
- ▶ Uniform definition of MTC across TG is welcome but not test limit - different species and endpoints/systems may have different sensitivities
- ▶ better explanation or instruction is desired
- ▶ see previous comments on MTC harmonization
- ▶ support the limit test level of 10 mg/L in line with similar assays. Some additional guidance around the MTC would also be welcomed and broadening the definition to include endpoints other than mortality would be supported.
- ▶ concentration of 100 mg/L is in most cases too high to be used in a flow through system
- ▶ It makes not sense not to be aligned with TG229

Disagreement justifications

- ▶ The current text in the guidance reflect the screening nature of the TG
- ▶ This should be discussed and agreed - if I remember well there is a rationale behind the choice of this limit concentration

6. para 16 - Change limit concentration to 10 mg/L



Agreement justifications:

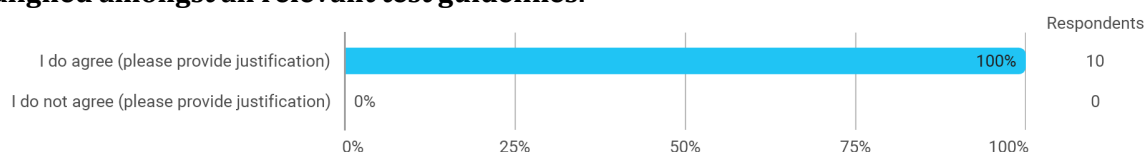
- ▶ 100 mg/L is a clear typo for a chronic study

- ▶ consistency with other TG
- ▶ support the limit test level of 10 mg/L in line with similar assays
- ▶ Agree. Alignment with fish TGs is desirable.

Disagreement justifications

- ▶ Maximum tested concentration should not be lowered if the test substance is still soluble and not systemically toxic at 100mg/L
- ▶ The current text in the guidance reflect the screening nature of the TG
- ▶ This should be discussed and agreed - if I remember well there is a rationale behind the choice of this limit concentration

7. para 17 - Test concentration setting guidance (The MTC) needs to be improved. It should be made clearer that mortality is not desirable. Suggest that MTC guidance is aligned amongst all relevant test guidelines.

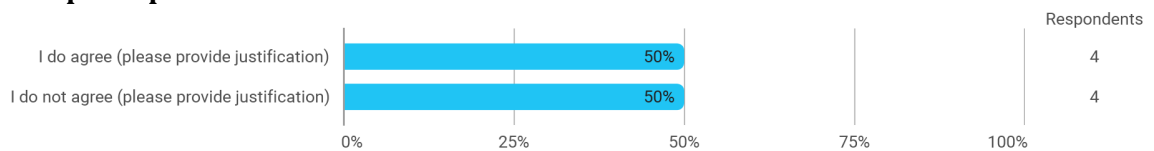


Agreement justifications:

- ▶ 100 mg/L is a clear typo for a chronic study
- ▶ I am not sure about this one. It is possible that testing only 2 concentrations is insufficient to cover the relevant activity spectrum, specially given the possible non monotonic effect of the endocrine systems
- ▶ agree
- ▶ better explanation or instruction is desired
- ▶ Some additional guidance around the MTC would be good so that we are not expected to kill tadpoles at the top level, but we need to maintain the criteria that the test is still valid if one of the test levels is lost due to systemic toxicity. It is not always possible to avoid any signs of systemic toxicity in the study design and there is pressure to test at high concentrations to make sure that ED is not missed by testing too low, so keeping the ability to drop the top dose and still have a valid study is helpful.
- ▶ Harmonization needed
- ▶ See comments on other TGs

8. para 30 - Suggested amendments in bold below: "Animals selected for histopathology (n= 20 tadpoles per treatment group from pooled replicates), should be matched to the median stage of the controls (pooled replicates) whenever possible." Rational: TG recommends stage matching to control median for thyroid histopathology. Thyroid histopathology is a qualitative assessment only. Therefore since pooled controls (pooling replicates) are used for obtaining control median, pooling replicates should be used for selection of stage matched individuals -

therefore recommendation should be for 20 tadpoles per treatment, rather than 5 tadpoles per replicate.



Agreement justifications:

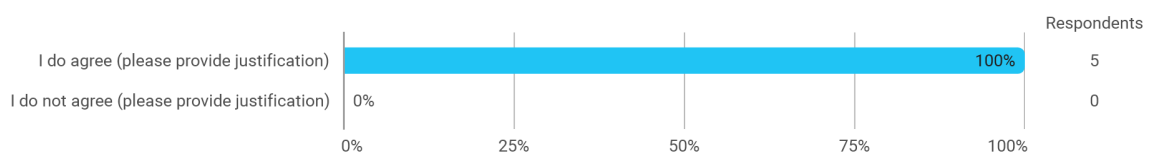
- ▶ To prevent differences between life stages
- ▶ Facilitates the selection of suitable stage matched samples and improve conditions for histopathological evaluation

Disagreement justifications

- ▶ Pooling replicates for histopathology greatly decreases the effectiveness of statistical analysis and clearly should not be done
- ▶ Not sure about this one either. It does have the advantage of maximizing the use of tadpoles available in a comparable developmental stage but may include undetected biases if the replicates vary a lot
- ▶ Tadpoles should not be pooled across tanks (i.e. treatments). When a developmental delay occurs, using the control median may not be the best option.
- ▶ I am not sure this would be of added value. It allows to take 10 tadpoles from 2 replicates rather than 5 from 4 replicates. It seems quite unbalanced to me.

9. para 31 - End of paragraph 31..."In the event that 5 tadpoles cannot be selected in the stage matched to the controls, additional matched stages should also be taken from the controls based on accelerated or delayed development"

Rational: Where 5 tadpoles matched to control median can not be selected from treatment groups, the TG recommends selecting additional tadpoles from one stage above for accelerated development or one stage below for delayed development. However there is not a great deal of clarity on how to proceed if 5 tadpoles can not be taken from the control median stage and the stage either below or above. Is the guidance to proceed down the stages (for delayed) and up the stages (for accelerated), or to straddle the control median. Since in this situation, there wouldn't be a comparison of like for like, it is recommended to also take additional matched stages from the control for better comparison.

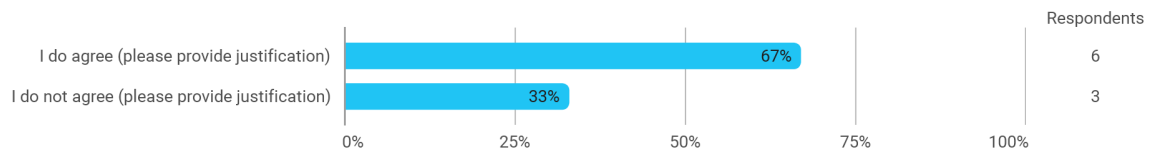


Agreement justifications:

- ▶ Can be a possibility to compare matching stages but might be difficult to implement
- ▶ Facilitates the selection of suitable stage matched samples and improve conditions for histopathological evaluation

- In general this practice is supported; however the implementation and analysis of the scenario is not necessarily straight-forward.

10. para 44 - Language offering more flexibility around the test performance criteria should be included.



Agreement justifications:

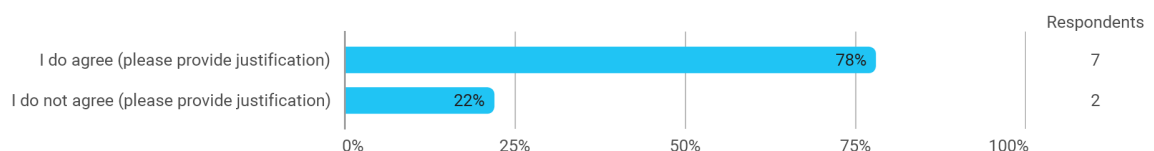
- allows the decision based on weight of evidence
- Specifications are sometimes misleading and not practicable
- It is not clear what flexibility is sought, but incorporating some flexibility around the test performance can be beneficial especially in regard to minimizing vertebrate animal use.
- these are often seen as validity criteria and they shouldn't be VC
- in some cases the validity criteria with negative thyroid activity is too strict

Disagreement justifications

- This is not a difficult study to achieve the performance criteria. No modifications are necessary.
- This TG is much clearer and still flexible than for instance FSTRA TG 229. I would prefer this one as a reference. more clear
- Test criteria should be met. By allowing too much flexibility we might end up having invalid test submitted.

11. para 45 - Reconsider validity criteria where the bar for a thyroid active substance is lower than that for a thyroid inactive substance. The uncertainties with the criteria should be the same in both cases.

Rational: Data quality should be the same to have confidence in either a positive or negative outcome.



Agreement justifications:

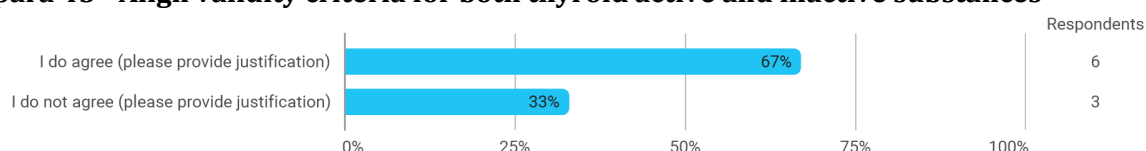
- allows the decision based on weight of evidence
- harmonization
- it is worth a discussion to consider aligning performance criteria for both thyroid active and in-active substances.

- ▶ Align validity criteria regardless of the outcome of the study (i.e., whether or not the test identifies the substance tested as active)
- ▶ agree seems strange that they are different

Disagreement justifications

- ▶ The observation of an effect proves it but a lack of observation does not disprove it, criteria must be more stringent. If an effect is observed the statistical power is less relevant than other wise. TG well designed
- ▶ I think this should be discussed. I think there was a rational beyond this choice when the guideline has been drafted. What has changed in our knowledge to change the validity criteria?

12. para 45 - Align validity criteria for both thyroid active and inactive substances



Agreement justifications:

- ▶ harmonization
- ▶ it is worth a discussion to consider aligning performance criteria for both thyroid active and in-active substances.

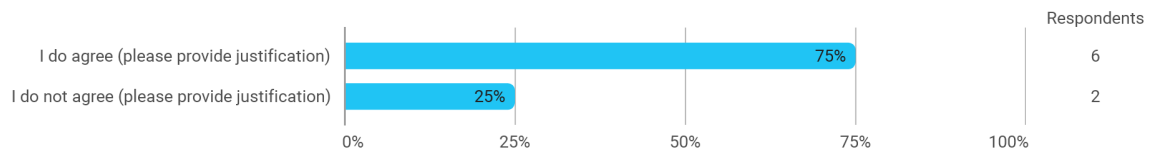
Disagreement justifications

- ▶ This does not allow the decision based on weight of evidence
- ▶ I think this should be discussed. I think there was a rational beyond this choice when the guideline has been drafted. What has changed in our knowledge to change the validity criteria?

13. para 51 - The recommendation is to revise the decision logic for thyroid agonists. The current decision logic is too conservation for a hazard-based cut off. Since the TG state that the weight of evidence is stronger if effects are observed on more than one of the four parameters, and/or more than one of the two timepoints, then a weight of evidence consideration of accelerated development based on effects on the four parameters at day 7 and day 21 should be considered in the assessment. Effects on one parameter at one timepoint for example, should not immediately be considered indicative for thyroid agonism.

Rational: The decision logic for accelerated development is too conservative for a hazard based cut off criteria. Effects on only one parameter at one time point is all that is required for the conclusion of accelerated development and effects at any time point do not need to be corroborated by effects on any of the other three parameters or timepoints. However the TG also states that the weight of evidence for accelerated development will increase if significant effects are detected for more than one endpoint. Therefore, if a weight of evidence is possible (and recommended), and conclusion of accelerated development can increase with effects on multiple endpoints (and timepoints), concluding accelerated development from effects on only

one parameter at one timepoint, for example, is not appropriate for the hazard-based cut off.



Agreement justifications:

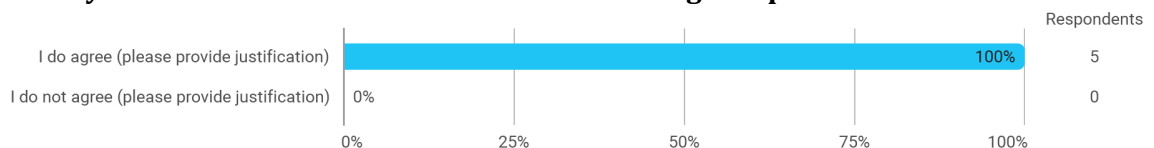
- ▶ I agree that the interpretation of positive effects on a single endpoint and sampling time deserves discussion
- ▶ Clearer guidance for data evaluation is desirable
- ▶ Day 7 can be a more sensitive timepoint for thyroid agonism since there is low baseline thyroid hormone levels at these earlier stages. However, lack of an effect carried through to day 21 would show potential for recovery and therefore I agree that a WoE approach looking at the complete set of endpoint responses at all timepoints is most scientifically defensible.
- ▶ OECD GD 150 recommends considering all parameters together (effect pattern approach), and not to conclude on isolated parameter responses.

Disagreement justifications

- ▶ Depending on the parameter even the indication of effects at one time point can provide information necessary for the positive detection of accelerated development. E.g. accelerated development on Day 7 when the tadpoles are functionally athyroid at the start of the test.
- ▶ I think this should be discussed. What has changed in our knowledge from when the guideline was published to change the decision tree?

14. para 72 - Suggestion to add recommendations for how to treat normalised hind limb length (nHLL) data for studies with >20% late stage tadpoles (>NF60).

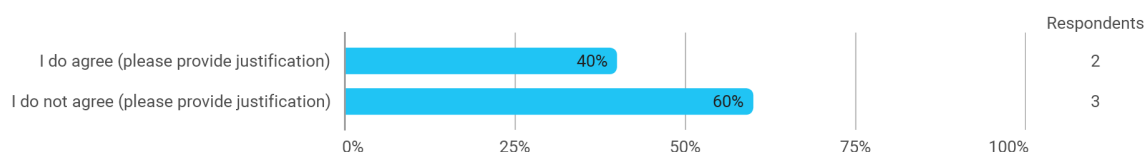
Rational: It is not clear in the current TG how normalised hind limb length (nHLL) should be treated in the case of >20% late stage tadpoles (>NF60). Should nHLL be treated similarly as snout-vent length and wet weight? Currently there is no consistency in how to treat nHLL data for >20% late stage tadpoles



Agreement justifications:

- ▶ recommendations help studies to become more uniform and therefore evaluation of different test items/ by different labs become more comparable and fair
- ▶ guidance for data computation and statistical evaluation is desirable
- ▶ There should be consideration to adding recommendations for how to treat normalized hind limb lengths data for studies with > 20% late stage tadpoles.

15. The methodology as described in the current guideline is not suitable to test for effects of virtually insoluble particulate materials (e.g. nano-forms of substances, polymers). Adaption of the guidelines methodology with latest scientific findings is indicated so that valid tests with particulate materials can be performed.



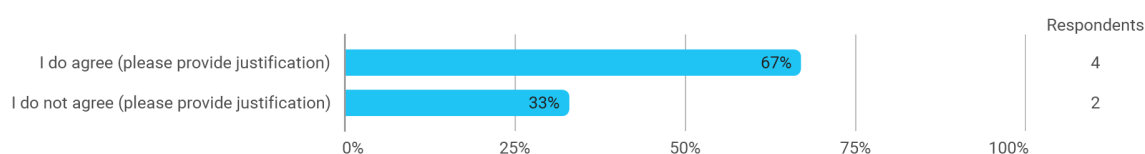
Agreement justifications:

- I have no expertise in insoluble particulate material - maybe some general guidance could be useful here

Disagreement justifications

- The issue should be addressed with in a higher-level guideline (OECD 23)
- Some general wording to allow for testing a diversity of materials could be included (i.e. for testing nanomaterials or polymers above water solubility limit etc.); however detailed instructions on testing difficult substances should not be addressed in individual toxicity testing guidelines, but should be preferentially addressed in guidance documents such as OECD 23.
- Methodology is not fully established yet and needs separate guidance.

16. Concerning hindlimb length data - There should be guidance added to the TG on the acceptable natural variability in hindlimb length data. It can be a variable endpoint, and acknowledgment and acceptability of natural variation should be outlined in the guidance in the Data and Reporting Section.



Agreement justifications:

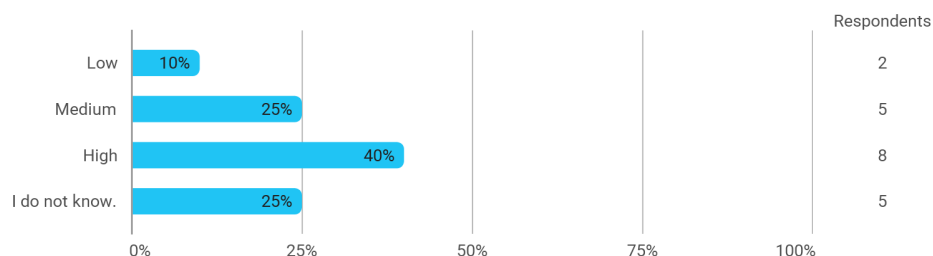
- the variability is also high for body weight and SVL
- I am not sure if a consensus on what "acceptable" would be; however some background information on the variability of the endpoint could be added.

Disagreement justifications

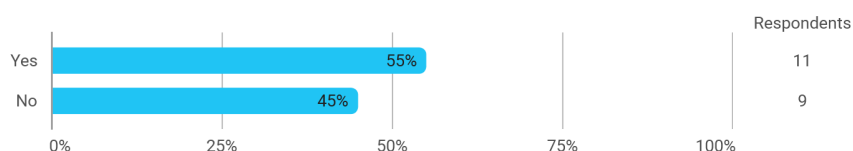
- I do not see the need, Animals come from the same batch and are distributed randomly
- While natural variation is useful, it is preferable to use historical control information from individual labs since husbandry may be a large factor in the variability.

C.24 Test No. 241: The Larval Amphibian Growth and Development Assay (LAGDA)

Relevance of revision

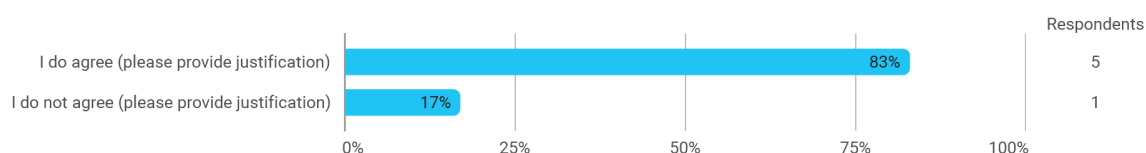


Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. **para 3 - The TG calls for a re-evaluation when there is sufficient experience with the test design. Is now the time to do this?**
Rational: The study is complex, animal intensive with stringent validity criteria. There may now be sufficient experience to allow for the test design and validity criteria to be re-discussed.



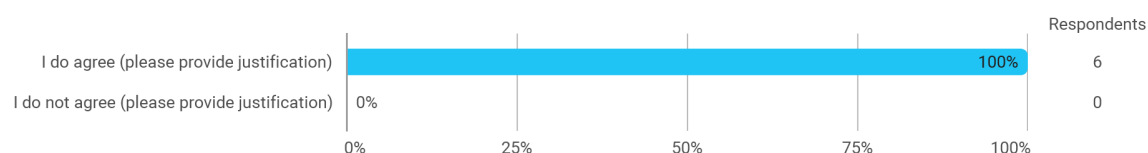
Agreement justifications:

- reevaluate current state of art
- Now this TG is in regular use this would be useful although I suspect there is insufficient data to do this (we have tried to collate relevant data for historical control data analysis)
- The LADGA raises serious concerns: test guideline not properly validated, test not specific for the thyroid modality, very animal intensive test, virtually no experience at the CRO level, high risk of test failure (need repetition), virtually no experience with the interpretation of the results.
- Strongly agree see MEOGRT comments

Disagreement justifications

- There is not adequate experience across CROs to make this a productive exercise at this time.
2. **para 8 - the scientific basis of the test method section of this TG only refers to the HPG axis, while it is understood that this test guideline provides information related to the**

thyroid modality (HPT axis). Consideration should be given to provide more information on the scientific basis for the test method.

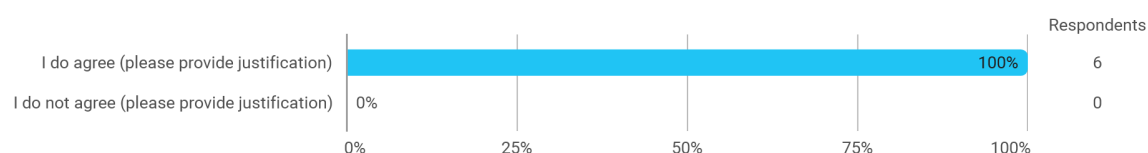


Agreement justifications:

- This TG goes beyond the scope needed
- It is fundamental to aggregate as much endpoints as necessary in as much TG as possible
- Agree, this issue has been highlighted against the TG when the study request has been contested in the regulatory context
- The HPT axis should be added to the scientific basis section; however, estrogen, androgen, and thyroid processes are considered characterized in this test according to paragraph 11.
- The domain of applicability of the LADGA should be clearly defined. The test is highly redundant with fish tests where more experience/expertise does exist.

3. para 25 - The MTC should be aligned with the other aquatic endocrine TGs.

Rational: To avoid regulatory uncertainty and the need for unnecessary repeat studies

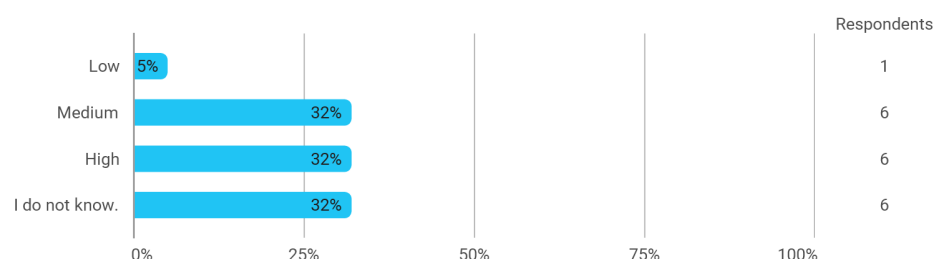


Agreement justifications:

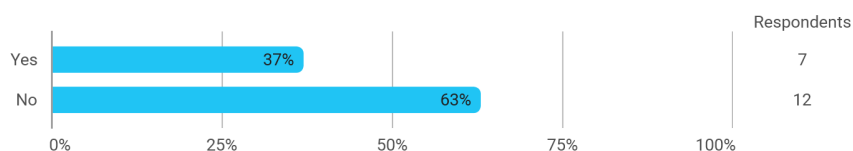
- Yes, but criteria for determining relevant concentrations should be refined - research on this topic required
- agree
- Alignment of the definition of the MTC across OECD TGs is desirable.
- see comments on other TGs

C.25 Test No. 248: Xenopus Eleutheroembryonic Thyroid Assay (XETA)

Relevance of revision



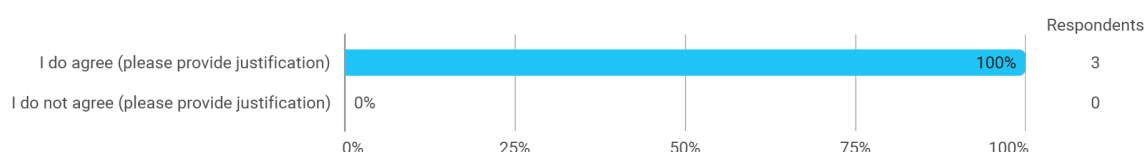
Interest in being part of an OECD Expert Group



The following suggestions have been collected:

1. paras 26, 27, 28 - The definition of the MTC should be consistent with the definition of the MTD (from OECD), and therefore include non-lethal effects, as also recommended by Hutchinson et al. (2009) and Wheeler et al. (2013). The MTC is determined prior to the definitive test (i.e. via calculation using existing ECx and NOEC values or from range-finding tests) and may not be observed in the definitive test due to biological variability.

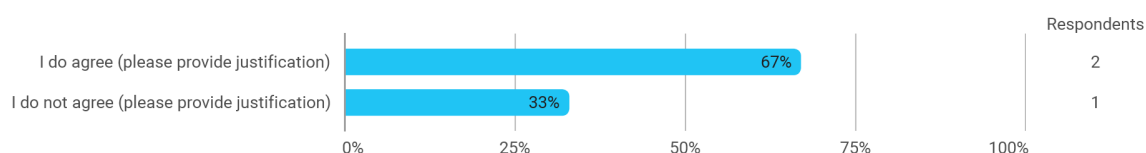
Rational: Consistency between OECD definitions of MTD and MTC.



Agreement justifications:

- The definition of the MTC should be aligned to that of the MTD and includes non lethal effects (e.g., reduced weight gain, behavioural/morphological effects, clinical signs of intoxication)
2. para 50 - A test chemical is considered to give a positive result in the XETA if two consecutive test concentrations including the highest are active in T3-spiked and/or unspiked mode. Whether it is observed in the definitive test or in range-finding studies, consideration of lethal and non-lethal toxicity should be included in the determination of active concentrations (especially the top concentration). The flow-chart in Figure 2 should be modified accordingly.

Rational: Consistency between OECD definitions of MTD and MTC.



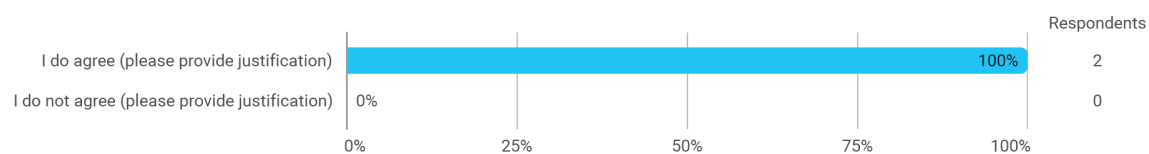
Agreement justifications:

- Conditions in which a test should be concluded as positive should be refined.

Disagreement justifications

- because a non-monotonic dose-response is not impossible for ED studies.

3. para 35 - Addition of control(s) and clarification of spiked unspiked mode



Agreement justifications:

- We can discuss/consider inclusion of a positive control