

TEXTE

193/2020

Necessary adaptations for a harmonized field- testing procedure and risk assessment of earthworms (terrestrial)

Summary

TEXTE 193/2020

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

Project No. (FKZ) 3715 67 420 0 Report

No. (UBA-FB) FB000314/ENG,ZUS

Necessary adaptations for a harmonized field-testing procedure and risk assessment of earthworms (terrestrial)

Summary

by

Jörg Römbke, Bernhard Förster, Stephan Jänsch, Florian
Kaiser, Adam Scheffczyk
ECT Oekotoxikologie GmbH, Flörsheim


Martina Roß-Nickoll, Benjamin Daniels, Richard Otter-
manns, Björn Scholz-Starke
RWTH Aachen University, Aachen


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:

ECT Oekotoxikologie GmbH
Böttgerstr. 2-14
65439 Flörsheim
Germany

RWTH Aachen University
Worringerweg 1
52074 Aachen
Germany

Report completed in:

June 2020

Edited by:

Section IV 1.3 Pesticides, Ecotoxicology and Environmental Risk Assessment
Silvia Pieper, Pia Kotschik und Susanne Walter-Rohde (Fachbegleitung)

Publication as pdf:

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

ISSN 1862-4804

Dessau-Roßlau, October 2020

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

Abstract: Necessary adaptations for a harmonized field-testing procedure and risk assessment of earthworms (terrestrial)

The purpose of this project was to provide scientifically robust and practical information on the variability of the endpoints assessed in earthworm field studies, the statistical significance of the results and the level of the statistically detectable effects of the chemicals tested - with the aim of developing suggestions for improving the test design. Best-practice studies reveal low power to detect differences between control and test chemical treatment plots. An adapted test design should contain an option to perform regression (EC_x) approaches, which have been suggested as an alternative to the currently performed threshold (NOEC) approach. A pilot field study was performed according to a newly developed combined NOEC- and EC_x-test design with the test chemical carbendazim. The EC_x design leads to more robust conclusions for environmental risk assessment. The calculation of effect thresholds (NOEC/LOEC) should be conducted with the most powerful multiple test procedure for given data prerequisites. If applicable to the data, the closure principle computational approach test (CPCAT) is the preferred option. The evaluation and interpretation of the data at plot (pooled samples of 1 m² in total used as replicates) and sub-plot level (single samples as replicates of 0.25 m²) should be requested. According to the experiences made during the performance of the pilot study and the results of the statistical analyses, a draft OECD test guideline was developed. As of now, the discussion of the draft test guideline is ongoing.

Kurzbeschreibung: Notwendige Anpassung zur harmonisierten Freiland-Testung und Risikobewertung für Regenwürmer (Terrestrik)

Ziel dieses Projekts war es, wissenschaftlich belastbare und praktische Informationen über die Variabilität der in Feldstudien mit Regenwürmern ermittelten Endpunkte, die statistische Signifikanz der Ergebnisse und die Höhe der sicher statistisch nachweisbaren Auswirkungen der getesteten Chemikalien zu liefern, um Vorschläge für die Verbesserung des Testdesigns zu entwickeln. Best-Practice-Studien zeigen, dass die statistische Trennschärfe zur Erkennung von Unterschieden zwischen Kontroll- und mit Testchemikalien behandelten Parzellen gering ist. Ein angepasstes Testdesign sollte eine Option zur Durchführung von Regressionsansätzen (EC_x) enthalten, die als Alternative zum NOEC-Ansatz vorgeschlagen wurden. Eine Pilotfeldstudie wurde nach einem neu entwickelten kombinierten NOEC- und EC_x-Testdesign mit der Testchemikalie Carbendazim durchgeführt. Das EC_x-Design führt zu belastbareren Aussagen für die Umweltrisikobewertung. Die Berechnung der Wirkungsschwellen (NOEC/LOEC) sollte unter den gegebenen Voraussetzungen mit dem leistungsstärksten Mehrfachtestverfahren durchgeführt werden. Wenn möglich, ist der CPCAT-Ansatz (closure principle computational approach test) die bevorzugte Option. Die Auswertung und Interpretation der Daten auf der Parzellen- (gepoolte Proben von insgesamt 1 m², die als Replikate verwendet wurden) sowie der Probenebene (einzelne Proben von 0,25 m² als Replikate) sollte gefordert werden. Basierend auf den Erfahrungen während der Durchführung der Pilotstudie und den Ergebnissen der statistischen Auswertungen wurde ein OECD-Prüfrichtlinienentwurf formuliert. Die Diskussion über den Prüfrichtlinienentwurf ist derzeit noch nicht abgeschlossen.

Summary

Introduction

Since 1994, the risk of chemicals for earthworms in the field is assessed by a test that was originally standardised by the German Federal Biological Institute (BBA). Since 1999, an international guideline standardised by International Organisation for Standardisation (ISO) is available that has been updated several times up to now (last in 2014) without changing the basic approach (ISO 11268-3). However, ISO guidelines focus on the assessment of (potentially) contaminated compartments (water bodies, sediments, waste materials as well as soils), i.e. they are used in a retrospective approach for an environmental risk assessment. In contrast, OECD test guidelines serve in general the purpose of a prospective assessment of individual chemicals and defined chemical mixtures such as pesticide formulations. As a consequence, several ISO guidelines used in the testing of chemicals were transcribed to the OECD format during the past 10 years. In the course of this conversion, which in the case of the earthworm field test is performed under German lead since April 2013 as OECD project no. 2.47 ('New Test Guideline on Determination of Effects on Earthworms in Field Studies'), it was also checked whether -apart from formal adjustments- further amendments were necessary. This assessment was performed by an ad hoc sub-group of the Global Soil Interest Group (GSIG) of the Society for Environmental Toxicology and Chemistry (SETAC) gathering representatives of academia, industry and authorities. Based on the experiences made during the past 20 years it was decided that several aspects of the guideline need adjustment to reflect the scientific progress. Specifically, besides technical details, the study design and the statistical evaluation of the test results had to be optimised. Regarding the study design, the ISO Guideline already mentions the possibility of performing studies according to a dose-response design, an option that is deemed to "clearly facilitate environmental risk assessment compared to single dose studies" (ISO 2014). In particular, due to the variability of the endpoints assessed in the field, the test design and evaluation needed improvement, so to increase the statistical significance of the results of the field test and the level of safely detectable effects of the tested chemicals. In addition, some assessment criteria proposed by the European Food Safety Authority (EFSA PPR 2017) needed to be translated in measurable endpoints. To address these issues, scientifically robust and practical information was missing. The generation of this information was the objective of this project. In close cooperation with the ad hoc SETAC GSIG sub-group, the following aims were reached by performing three work packages (WP):

- ▶ WP1: Evaluation of existing data and development of proposals for an optimized design of the earthworm field test: Compilation and critical evaluation of information available in the literature and the database of the German Environment Agency (UBA) regarding the standardised performance of earthworm field studies to develop an improved test design;
- ▶ WP2: Experimental investigations and statistical analyses: (1) Performance of a pilot field study according to the new test design. (2) In-depth statistical analysis of the pilot field study in combination with the existing database regarding natural variability in earthworm communities. (3) Calculation of effect thresholds, effect concentrations and community analysis. (4) Formulation of design requirements for earthworm field studies and identification of limitations and open questions;
- ▶ WP3: Participation in the OECD process: Formulation of a new draft OECD test guideline (TG) based on the existing ISO guideline 11268-3 but following the formal requirements of

the OECD, using the experiences made in the pilot study as well as the evaluation of the UBA database. Discussion of this draft guideline within the ad hoc SETAC GSIG sub-group in a final project meeting. The combined results of the development and discussion process will be submitted to OECD.

Evaluation of existing data and development of proposals for an optimized design of the earthworm field test (WP 1)

In the course of the preliminary analyses and investigations, the ISIS database (“Information System Chemical Safety”) of the UBA was identified as a useful source for data analysis of earthworm field tests. The database held 150 entries for field studies on earthworms. Quality criteria for data were initially defined with regard to further statistical investigations. Raw data “abundance” and “biomass” on sample level (0.25 m²) were extracted from original study reports. A unified database was developed for further statistical analysis. The subsequent systematic procedures of descriptive metadata analysis and advanced statistical calculations were performed.

Earthworm field study database – compilation and quality check

Only earthworm field studies possessing the following characteristics were used for statistical analyses: Earthworms should have been sampled by a combination of formalin/allyl isothiocyanate (AITC) extraction and hand-sorting. A bias of the sampled species composition due to the use of the octet sampling was therefore prevented. Moreover, the technical reports should include raw data collected on the sample (= subplot) level. This prerequisite enabled an analysis of test data at sample level in comparison to the conventional evaluation at plot level. The 21 field studies that fulfilled these characteristics were divided into two classes: Tests with only one treatment and one reference compared to the control (limit test) were assigned to class 1, while tests with several treatment levels were classified as class 2. Eleven field studies were classified into class 1 (limit-tests), two field studies assessed two different substance concentrations next to the control, and another eight field studies were designed with three treatments (class 2). In addition, further 5 studies with digitalized raw data at sample or plot level were integrated into the database, each with a slightly different sampling method. In total, data of 26 field tests of the ISIS database (+test data of the pilot study performed in this project) were used for statistical calculations. The processed field studies were carried out according to the ISO guideline 11268-3 or in consideration of the BBA (Biologische Bundesanstalt) guideline part VI, 2-3. Therefore, the analyzed test procedures followed a common approach. All reports contained information on earthworm species, numbers, and biomass collected for sampling plots treated with a test substance in a randomized arrangement (four replicates per treatment) and compared with those collected from control and reference plots. Every replicate (=sampling plot) consisted of four aggregated samples (= subplots) of 0.25 m² per sample (1 m² sampling plot in total). The sampling dates were usually set about 1-3 months, 4-6 months and 12 months after application. Tests usually started in April/May. The calculations of effects within the test procedures were mainly limited to the evaluation of abundance and biomass on species level and for total earthworms. Juvenile earthworms were summarized and evaluated on genus level (morphological groups: *tanylobous* and *epilobous*). In addition, the ecological groups of endogeic, epigeic and anecic earthworms were differentiated.

Data collection: environmental and agricultural variables

Descriptive metadata of the field studies revealed that the composition of species among all field studies consisted of 6 to 14 species per study. The respective Shannon Diversity Index was between 0.3 and 1.6 (mean: 1.2). The diversity index was slightly higher on grassland sites (mean: 1.44) than on other land use types (bare soil: 1.27; crop sites: 1.05). Accordingly, the minimum

number of species in grassland was at least 10. The mean number of individuals sampled was about 372 per m² on grassland, 356 on bare soil and about 196 on crop sites. The dataset available did not allow for an in-depth analysis of the potential systematic impact of environmental conditions or land use type on the earthworm community.

Field study data: Species composition, variability and MDDs

Based on the ISIS-database pre-processing, data of field studies for earthworm communities were subsequently analysed. The sampled individuals of the 21 field studies belonged to 17 different species. As a statistical measure, the minimum detectable difference (% MDD) between control and treatment of all field studies was calculated. Although the most likely value of the MDD for abundance data of total earthworms in the database was 45%, the probability of obtaining an MDD smaller than 50% of the control was 42%. An MDD between 10% and 35% (proposed in the EFSA soil opinion (EFSA PPR 2017) as small effects on the protection goals) was calculated with a probability of 8%. The same calculations for total biomass gave even lower power values than for total abundance: an MDD smaller than 50% was only detected for 32% of all sampling time points. For the aggregated group of total earthworms, the most powerful MDDs were calculated. For the most dominant species in the database, *Aporrectodea caliginosa*, the possibility to detect statistically significant effects in the field studies was even worse. Individuals of *A. caliginosa* had a very low probability to show MDDs less than 50% (12% of all sampling time points within the database). The most likely value of the calculated probability distribution for MDDs of *A. caliginosa* was 66%. Again, even lower MDDs were calculated for the endpoint biomass. In an overall picture, best-practice studies (using a combination of hand-sorting and formalin/AITC extraction for earthworm sampling) revealed low power to detect differences between control and treatment plots for aggregated taxa. Thus, based on statistical considerations, the testing and adaption of a new field study test design in the course of this project was justified. The limitations of the old design, covering limit-tests as well as NOEC-approaches, became evident. Therefore, an adapted test design should contain an option to perform regression approaches as an alternative to the NOEC approach.

Development of a pilot study test design

In a joint discussion between the UBA and the project consortium, the results of the evaluation described above led to a first proposal of the earthworm pilot field study design to be performed in 2017. This design of the experimental pilot study was characterized by combining a so-called NOEC- with an ECx-design and was called “mixed omni-design”:

- ▶ Four sampling dates, covering a total test duration of one year (as in ISO guideline 11268-3);
- ▶ One control (C) and six test chemical treatments (T) (only limit test in the ISO guideline);
- ▶ Number of plots per treatment six (C, T2, T5) or three (T1, T3, T4, T6) (four in the ISO guideline);
- ▶ Five samples per plot (four in the ISO guideline).

Running such a study meant that in total 30 plots with 150 samples per sampling date had to be covered. This original proposal was considered by the project team as large but still practical in terms of handling (e.g. number of days needed for sampling, field size etc.).

This proposal of the test design for the pilot study was discussed during the meeting of the ad hoc SETAC GSIG sub-group in February 2017. Further recent contributions addressing different aspects of the planning, performance or evaluation of earthworm field studies were presented to

the group. In the following discussion during the meeting various changes to the “mixed omnidesign” were proposed, all of them with the intention to improve the quality of the study output but without strongly increasing the efforts at the same time. The resulting final test design for the pilot study was called “balanced design”. It was decided to take six samples per plot in the NOEC- as well as in the ECx-plots and the number of replicate NOEC- and ECx-plots were six and three per treatment, respectively.

The selected test chemical was carbendazim, since it is by far the best-studied pesticide in soil ecotoxicology due to its use as reference substance in earthworm laboratory and field tests. Using the available information, various carbendazim concentration ranges were discussed. The following six application rates (plus a negative, i.e. water-only, control) were finally selected to cover a range spanning from concentrations where no effects are expected to concentrations where strong effects are likely: 0.6, 1.8, 3.2, 5.8, 10.5, and 31.5 kg carbendazim/ha. In the currently used ISO guideline 11268-3, the reference substance carbendazim should yield a statistically significant difference of at least 50 % on overall abundance and/or biomass compared to the control at least at one sampling date, when applied at rates of 6 to 10 kg a.s. carbendazim/ha. Thus, such effects should be detectable at the three highest application rates. Accordingly, and referring to the experiences made in an EU project focusing on the development of a standard semi-field method where Terrestrial Model Ecosystems (TME) have been employed, no detectable effects should appear at the two lower rates. A priori analyses have shown that an EC₅₀ could be expected at rates around 2.5 kg carbendazim/ha.

Experimental investigations and statistical analyses (WP 2)

Performance of the pilot field study

Arable land was chosen for the trial. It was surrounded by agricultural fields and pathways. The experimental plots were installed within an area of approximately 55 m by 107 m. Winter wheat was grown on the field before the study took place. To free the experimental site from vegetation without soil tillage that would have impacted the earthworm community, glyphosate was applied at a rate of 1.8 kg a.s./ha. For each treatment, i.e. control (C) and six different test chemical (carbendazim) treatments (T1 to T6), six (C, T2, T5) or three (T1, T3, T4, T6) plots (= replicates), each 10 m by 10 m, were installed at the field site and assigned randomly. The distance between two neighbouring plots was 3 m and the distance to the surrounding fields or cart tracks was at least 5 m. The test chemical was applied as the suspensible concentrate (SC) formulation Carbomax 500 SC once on 11 April 2017. The water (control) and the test chemical were applied onto the bare soil surface at a wind velocity below 3 m/sec to avoid any risk of cross contamination due to possible drift during application. All experimental plots were irrigated directly after application by means of a tractor-pulled tank wagon with at least 1000 l/plot (equivalent to 10 mm precipitation). The experimental plots were left to natural development of vegetation. No agricultural practices such as tillage, application of plant protection products or fertilizers, were undertaken. On 25 August 2017 all plots were mowed with a string trimmer and all cuttings were left on the plots.

Eight to six days prior to the first application of the test chemical, earthworms were sampled on all plots. The mean total number and the mean biomass of earthworms were determined for each of the thirty plots, designated either for test chemical treatment or to serve as untreated controls. The mean number of earthworms collected (hand-sorting and AITC-extraction) before application ranged from 413 to 512 ind./m² - hence fulfilling the requirements of the ISO guideline 11268-3. Earthworms were sampled at each sampling time point by a combined hand-sorting and AITC extraction method. Six random samples of an area of 0.25 m² (50 cm x 50 cm) to a depth of approximately 20 cm were taken per plot. Hence, there were 18 (3 plot replicates)

or 36 (6 plot replicates) individual samples per treatment and sampling time point. The distance between two samples taken on the same date and plot was at least 2 m. The sampled area was marked and not used again at subsequent sampling dates. Samples were taken at least 2 m apart from the plot border. Five to ten litres of an AITC solution (0.1 g/l) were poured uniformly into the remaining cavity to catch earthworms from deeper soil layers. The soil was carefully searched for earthworms by hand-sorting. These worms and those extracted by AITC were preserved in a 70% ethanol solution in watertight containers.

The worms were identified by means of a binocular microscope, using morphological characters. Adult worms were determined to the species level. Juveniles were classified according to the genus level, but in some cases a distinction of small worms belonging to closely related genera was not possible (e.g. *Allolobophora* and *Aporrectodea* were combined). All adult worms of one sample belonging to a particular species and all juvenile worms belonging to a particular genus were weighed together. The field site was inhabited by an earthworm population which can be considered typical for central European arable land (ISO 11268-3) including the ecological most important groups of anecic and endogeic earthworms. In total, nine different species of earthworms were found during the study. The lumbricid biocoenosis was dominated by juveniles of the endogeic genera *Aporrectodea/Allolobophora* and *Allolobophora chlorotica* was the most abundant species.

The test chemical Carbomax 500 SC (a.s. carbendazim) caused a clear reduction in total abundance and biomass at all three post-application sampling time points. Compared to the control, mean abundance and mean biomass in the test chemical treated plots were 15-59% and 11-55%, respectively at 34-36 days after application (DAA), 45-90% and 69-111%, respectively at 188-190 DAA, and 38-74% and 80-113% respectively at 377-379 DAA.

Statistical analysis: field study and database

A set of different statistical data analysis procedures were conducted for both data of the pilot study and existing test data from the UBA database. The main focus was to improve the conventional statistical methods to evaluate earthworm field studies (ISO 11268-3) and to acquire insights for statistical considerations regarding an adapted test design for earthworm field studies. Raw data for biomass and abundance at all sampling dates and for all taxa and morphological or functional earthworm groups were integrated at sample- and plot level into the existing database of the project. Single species calculations and analyses did not show any significant effects at the last sampling time point (377-379 DAA). With “*Aporrectodea/Allolobophora* spp. juvenile”, a statistically significant effect could be observed in a taxonomic group after one year. Due to the high dominance of this group in the overall data set, a reduction of abundance and biomass after 12 months was also indicated in other aggregated groups. However, this was exclusively caused by juveniles of *Aporrectodea/Allolobophora* spp. This example illustrates the need for assessments of different types of endpoints and earthworm groups (e.g. species level and group level), to avoid only general conclusions for effects of test substances based on an aggregated endpoint such as total abundance of all earthworms.

The natural, heterogeneous scattering of earthworm species within a field is a decisive factor for the statistical visibility of possible effects caused by applied environmental chemicals. The variability of tested endpoints in database and pilot field studies was assessed using the coefficient of variation (CV) of field study control treatments to derive conclusions and suggestions for improvement regarding the test power. The natural variability of the species groups in field studies was illustrated descriptively as the variance of the control treatments and used as a basis for multiple sample planning. Aggregated earthworm groups had the lowest CVs while rare species showed a comparatively high relative scattering between plots. Results indicated that particularly

aggregated species groups with high abundances and biomass values provide powerful end-points (especially low variation in controls and treatments). On average, the scattering at the single species level seemed for many species too high to prove statistically significant effects. A high variation in control treatments thus leads to a lower detectability of possible effects of the test substance (= high MDD).

The impact of variance on the number of required replicates to achieve a certain test power was determined for the standardised Dunnett test. Calculations were based on CVs for control treatments in earthworm field tests and applied for a dynamical sample size planning for an MDD that should be achieved. For the development of an adapted test design we investigated how many samples (=replicates) should be used given a desired target-test power and a given natural variability of data. By default, the desired test power is usually set to 80% for statistical hypothesis testing. The MDD that can be achieved with the respective sample sizes was classified into four different classes in the simulation, adapted to the scaling of magnitude of effects on the protection goals as proposed in the EFSA soil opinion¹. Up to 10% difference between control and treatments was defined as negligible deviation, up to 35% as small effects, between 35 and 65% as medium effects and higher than 65% as large effects. Even if it is not required to measure these effect ranges in field tests, a comparison with the available data was performed. Results of the sample size simulation for mean total earthworm variability indicated that standardized earthworm field tests might have an insufficient number of replicates to detect small effects with a test power of 80% for “total earthworms”, which was the group with the lowest CVs in earthworm field test data. Accordingly, the ability to detect effects for other earthworm groups is even more limited.

For this reason, it was investigated if a NOEC calculation using samples (=subplots) as statistical replicates would result in an improvement with regard to MDDs. We calculated the sample size planning with the measured CVs on plot level (current standard method), and on sample level to assess shifts in test power. Results of the pilot study showed that the increase in plot numbers (n=6) and the slightly lower CVs (1.5 m² sampled instead of 1.0 m²) increased the test power. The number of required replicates to achieve a certain threshold of MDDs decreased compared to database field studies. Nevertheless, using this test design, medium effects (35% - 65% effect) will very often not be detectable with a power of 80%. The comprehensive detection of small effects (10% - 35%) with a test power of 80% appears not to be achievable in this simulation considering realistic numbers of replicates. Nevertheless, a 35% difference would be detectable at sample (= subplot) level. In this case, the mean CV at subplot level would be slightly higher than at plot level (34.56%). For this reason, at least 14 replicates would be necessary to detect medium effects. By using the single samples as replicates there would be 36 replicates available in this statistical design. This switch in the assessment level allows the identification of more significant differences, especially at single species level. In general, the statistical detectability of effects always improves if the evaluation is carried out at the subplot level.

The calculation of the effect thresholds was carried out for all studies in the database and at all sampling times and a comparison of the Dunnett method with results from the so-called CPCAT approach (Closure Principle Computational Approach test) was performed. The theoretical distribution assumption of earthworm abundance field test data follows a Poisson model. Therefore, the application of the CPCAT approach is highly recommended for abundance data due to more powerful test statistics. This is the first time in which the performance of CPCAT was assessed within a comprehensive meta-analysis of field study data. It was shown that the use of the CPCAT

¹ EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues) (2017): Scientific Opinion addressing the state of the science on risk assessment of plant protection products for in-soil organisms. EFSA Journal 15(2):4690, 225 pp. doi: 10.2903/j.efsa.2017.4690

procedure in comparison to the Dunnett test increased the probability that significant effects were identified, even at small effects (10% - 35%). CPCAT is therefore generally "more selective", i.e. significant effects of the test substance are already indicated for smaller differences to the control. The differences in test power between the two procedures also became visible in the separate examination of the NOEC calculations for different species groups. The Poisson distribution used in CPCAT procedures describes the earthworm community data in outdoor tests mathematically and statistically more accurately than the normal distribution used in conventional t-tests (e.g. Dunnett or Williams). Thus, the use of the CPCAT approach increases the test power for earthworm field data. However, for the relatively new CPCAT approach there is currently no procedure for the generic calculation of a quantitative measure of test power and sample planning using the CPCAT procedure is not yet possible.

In contrast to the database studies, the pilot study was carried out in a test design with several concentration levels. Probit curve regressions were conducted and at all three samplings after application, a significant dose-response relationship was identified in the group "total earthworms". Unlike the NOEC approach (Dunnett test), the choice of an EC_x design allowed revealing significant relationships between the carbendazim concentration applied and the measured effect on the earthworm population (total abundance in the pilot study) across the whole range of concentrations. In addition, the comparison of EC₅₀ values across sampling times indicated recovery effects that could be assumed in case of the total earthworm community between 1 and 6 months after application. The calculated EC₅₀ increased by a factor of 10 during this period, whereas it did not change in any further comparison at the sampling time after 12 months. By contrast, the EC₅₀ for (*Aporrectodea/Allolobophora* spp.) juveniles increased only by a factor of approximately 3 until the end of the study. The results of the study showed that the use of an EC_x design to derive effect concentrations on the earthworm population in the field is generally feasible. However, the choice of a suitable concentration range for adequate testing of all species and aggregated groups poses a challenge.

A Principal Response Curve (PRC) was used to answer the question whether there was a significant relation between community structure and treatments. The PRCs revealed a highly significant effect of the treatment on the earthworm community (p-value < 0.05). A clear dose-response relationship was visible and with increasing concentrations the deviation from the control increased. According to the PRC, a recovery of the community (abundance of adults for single species regaining initial state) could be assumed at all test concentrations after approximately one year.

Based on these investigations, generic derivations of recommendations are limited due to the high variability in the various earthworm data sets of different field tests and due to the expectable impact of local site conditions. However, the following basic recommendations and requirements regarding the implementation and evaluation of earthworm field tests were identified:

1. There is still a need to determine and evaluate biomass and abundance at species level, as the aggregated morphological or functional groups used may disguise effects on single species.
2. The EC_x design is a meaningful alternative to the NOEC design in the earthworm field test. At least a mix design would be advisable. In fact, the EC_x design leads to stronger/more protective statements for environmental risk assessment (ERA) especially at lower effect ranges, a masking of possible effects as in the NOEC evaluation is avoided.
3. The calculation of effect thresholds (NOEC/LOEC) should be conducted with the most powerful multiple test procedure for given prerequisites. If possible, the CPCAT approach is preferred. If data are metric (e.g. biomass), multiple t-test procedures such as Dunnett's or Wil-

Williams' test ($\alpha = 0.05$, two-sided for unclear direction of response) should be performed for multiple comparisons in a randomized plot design. The prerequisite of normally distributed data and variance homogeneity has to be tested using e.g. Shapiro-Wilks and Levene's test procedures, respectively. If data do not fulfil the criterion of normality, generalized linear models or non-parametric tests e. g. the Bonferroni U-test or the Jonckheere-Terpstra Step-down-test (homogeneity of variance required) can be applied. The theoretical distribution assumption of earthworm abundance field test data follows a Poisson model. Therefore, the application of the CPCAT approach is highly recommended for abundance count data due to more powerful test statistics. Nevertheless, if abundance data show homogeneity of variances, the null-hypothesis of normal distribution is not rejected and absolute abundances per replicate are > 5 , the application of parametric test procedures (Williams, Dunnett) is also feasible. For multiple t-test procedures and with unequal replication, the table t-values must be corrected as suggested by Dunnett and Williams. In addition, an inappropriate log-transformation of data during the calculation procedure should be avoided.

4. After data revision, it should be decided whether a simple two-parameter Probit (Logit, Weibull) regression, a nonlinear regression or the integration of a so-called hormesis model for the calculation of effect concentrations (EC_x) is necessary. In case of a monotonous increase of the measured endpoint (biomass, abundance), the derivation of significant effect concentrations should also be taken into account.
5. If there are no ecological reasons for not using the data at sample level (i.e., no proven interdependency between samples from the same plot), the evaluation and interpretation of the data at plot (pooled samples of 1 m² in total used as replicates) and sub-plot level (single samples as replicates of 0.25 m²) should be requested.
6. Principal response curves are generally applicable within the EC_x -design and a powerful tool for community analyses. They should be carried out in addition to uni-variate methods when appropriate data are available, for tests with multiple treatments (e.g. EC_x design).

Some limitations and open questions regarding the proposed changes need to be kept in mind. The recommendations towards adjustments of the field study test design reveal two opposing trends whose benefits and downsides for the significance of the test have to be balanced. On the one hand, as many test-concentration levels as possible should be considered for a meaningful EC_x design. From a strictly statistical point of view, replication of the concentration levels is not needed for the subsequent regression analysis. A strong design for calculating robust NOEC values requires, as shown, a substantial increase in the number of replicates per control and treatments. These two demands need to be weighed and integrated into a new design depending on the underlying test concept and desired endpoints. However, this decision is not a strictly statistical one, but primarily a question of feasibility in the field (plot numbers and field sizes to be handled) and a question of regulatory prioritization of various endpoints.

In addition, the analyses and underlying data presented above have a few limitations that should not be forgotten: The results for the implementation of an EC_x design in field studies are based on a proof-of concept pilot field study at one site and with the well-known reference substance carbendazim. In this case, a sound prior knowledge and experience from earlier field studies on possible effect widths and dynamics was available. This is not the case, in particular, for new substances in regulatory practice. In such cases, the choice of concentration ranges in earthworm field tests might be considerably more difficult. Furthermore, the pilot field study demonstrates that an applied concentration range provides different dose-response curves for earthworm species and groups due to their different sensitivities, as also in the previously used NOEC design. If some species do not react to the test chemical, then no dose-responses and NOEC can be derived. However, the statistical endpoint of the NOEC disguises this to a large extent.

For the derivation of NOEC values with abundance data, CPCAT represents a meaningful alternative to the standardized test procedures of t-test statistics. However, it should be mentioned that there is still no established methodology for the calculation of test power and corresponding sample planning for CPCAT. Also, CPCAT should achieve higher acceptance as an appropriate tool for assessing the results of ecotoxicological tests, for example by being applied as a standard analysis method in a wider range of standard ecotoxicological test methods.

The CPCAT procedure is not suitable for metric data because the Poisson distribution does not adequately describe this type of data. To improve the statistical test procedures for metric data, it might be considered to integrate the closure principle into multiple t-test procedures to prevent alpha inflation.

The use of the samples as replicates for the calculation of NOEC values leads to an improvement of the test power. A general investigation of the effects in earthworm field tests at both plot and sample (= subplot) level is therefore recommended based on these results (provided that ecological conditions exist for the use of subplots as replicates). Whether this is a useful option in consideration of the debate on pseudoreplicates in field studies remains to be discussed. Within a regulatory framework, the following steps could be considered: A respective endpoint is evaluated at both subplot and plot level. If the same NOEC values are obtained as results, these are considered; if other (smaller) NOEC values are calculated at subplot level, the following procedure is suggested: If it is not possible to reliably demonstrate a relic of the plot effect at this level, the smaller NOEC should be used for the regulatory process. This is not necessarily a decision based on purely scientific considerations, but a regulatory, protective decision based on the precautionary principle.

Derivation of a new test design

The experience gained during the performance of the pilot study as well as the statistical evaluation of this study and the UBA database were applied to derive a proposal for a new test design (Table 1).

Table 1: Number of plots and treatments for the ECx- and the mixed-design in earthworm field tests. More information on the design type in the text above. C control; T 1-x treatments; R reference substance

Test design	Plots per treatment (No.)									Plots (sum)	Samples (total No.)
	C	T1	T2	T3	T4	T5	T6	(T7)	R		
ECx Design	3	3	3	3	3	3	3	(3)	3	24 (27)	96 (108)
Mixed Design	6	2	6	2	2	6			3	27	108

Participation in the OECD process (WP 3)

The experience gained in the more than 20 past years of performing earthworm field studies based on the existing BBA and ISO guidelines and during the project was used to formulate a new draft OECD test guideline including a proposal for a new test design. The draft OECD test guideline was distributed to the ad hoc SETAC GSIG sub-group in March 2019 as a basis for discussion during the final project meeting at UBA in Dessau. A multitude of comments were provided during and after the meeting which were compiled in a commenting table according to the

OECD process. This table is currently under review to create an updated version of the draft test guideline that will then be subject to the further OECD process.

Conclusions and outlook

The purpose of this project was to provide scientifically robust and practical information on (1) the variability of the endpoints assessed in earthworm field studies, (2) the statistical evaluation of the results and (3) the level of the statistically detectable effects of the chemicals tested. The final aim was to provide suggestions for an improved test design. Critical evaluation of information available in the literature and the database of the UBA revealed the following shortcomings of the currently used earthworm field test design according to ISO standard 11268-3:

- ▶ The evaluated best-practice studies (i.e. using a combination of hand-sorting and formalin/AITC extraction) reveal low statistical power to detect differences between control and treatment plots for aggregated taxa. For single species, this statistical potential for a reliable identification of effects is even lower. The overall MDD is not low enough for a comprehensive detection of small or medium effects.
- ▶ NOEC and related concepts have long been criticized in the ecotoxicological literature. Furthermore, the actual MDD calculations of field studies revealed that potentially relevant effects are not detectable in many field situations by the current standardized statistical procedures.
- ▶ An adapted test design should contain an option to perform regression analyses, which have been suggested as an addition to the NOEC approach. The resulting estimated concentrations (EC_x values) from fitting a curve to the data have been proposed as a more meaningful alternative to the NOEC-value. Thus, the number of concentration levels in the pilot field study has to be increased to investigate the suitability of an EC_x-design for earthworm field studies.
- ▶ To still include the possibility of deriving NOEC values as well as to improve the statistical power of this procedure compared to the old design, the number of replicates on the plot level for the control and test concentration treatments need to be increased.
- ▶ The number of samples per replicate should be increased to examine the changes in variance and to estimate if these samples can be used as individual replicates to improve statistical test power.
- ▶ As the field conditions and practical feasibility of the pilot field study limited the total number of plots, the enlargement of the concentration levels and the increase of plots and samples (=subplots) per treatment had to be adjusted in such a way that both research questions (feasibility of EC_x design and improvement of NOEC design) could be addressed.
- ▶ Based on these evaluations, a pilot field study was performed according to a newly developed combined NOEC- and EC_x-test design with the test chemical carbendazim. One control (C) and six treatments (T) were used. The number of plots per treatment were six (C, T2, T5) or three (T1, T3, T4, T6). The number of samples per plot was six. The results of the pilot field study and the in-depth statistical evaluation of additional earthworm field studies yielded the following design requirements for earthworm field studies:

- ▶ Abundance and biomass should be determined and evaluated at species level as aggregated morphological or functional groups may disguise effects on single species.
- ▶ The ECx design is a meaningful alternative to the NOEC design but at least a mixed design would be advisable. The ECx design leads to more robust conclusions for ERA, a masking of possible effects as in the NOEC evaluation is avoided.
- ▶ The calculation of additional effect thresholds (NOEC/LOEC) should be conducted with the most powerful multiple test procedure for given prerequisites. If possible, the CPCAT approach is preferred.
- ▶ If there are no ecological reasons for not using the data at sample level, the evaluation and interpretation of the data at plot level (pooled samples of 1 m² in total used as replicates) and sub-plot level (single samples as replicates of 0.25 m²) should be requested.
- ▶ Principal response curves are generally applicable within the ECx-design and a powerful tool for community analyses. They should be carried out in addition to uni-variate methods when appropriate data are available, i.e. for tests with multiple treatments (e.g. ECx design).

Some limitations and open questions regarding the proposed changes need to be kept in mind:

- ▶ There are two opposing trends whose benefits and downsides for the significance of the test have to be balanced: On the one hand, as many concentration levels as possible should be considered for a meaningful ECx design (with no replication of concentration levels required) while on the other hand a strong design for calculating robust NOEC values requires a substantial increase in the number of replicates per control and each treatment. This question is not a strictly statistical one, but it is also related to the feasibility in the field (plot number and field size) and of the regulatory prioritization of statistical endpoints;
- ▶ The results for the implementation of an ECx design in field studies are based on a proof-of-concept pilot field study at one site and with the well-known reference substance carbendazim. For new chemicals, the choice of concentration ranges might be considerably more difficult;
- ▶ There is still no established methodology for the calculation of test power and corresponding sample planning for CPCAT;
- ▶ The CPCAT procedure is not suitable for metric data because the Poisson distribution does not adequately describe this type of data. To improve the statistical test procedures for metric data, it might be considered to integrate the closure principle into multiple t-test procedures to prevent alpha inflation;
- ▶ The use of samples as replicates for the calculation of NOEC values leads to an improvement of the test power. A general investigation of the effects in earthworm field tests at both plot and sample (= subplot) level could therefore be recommended (provided that ecological conditions exist for the use of subplots as replicates). This is not necessarily a decision based on scientific principles, but a regulatory, protective decision based on the precautionary principle.

According to the experiences made in the more than 20 past years of performing earthworm field studies based on the existing BBA and ISO guidelines and during the project, a draft OECD test guideline (TG) was formulated and provided to the ad hoc SETAC GSIG sub-group for discussion. As of now, the discussion of the draft TG is ongoing.