Organisation for Economic Co-operation and Development

ENV/CBC/MONO(2026)8

Unclassified **English - Or. English**

7 July 2026

ENVIRONMENT DIRECTORATE

CHEMICALS AND BIOTECHNOLOGY COMMITTEE

Validation report for Test Guideline No. 256 on determination of effects on earthworms (Oligochaeta, Annelida) in field studies

OECD Series on Testing and Assessment

E-mail: ehscont@oecd.org

JT03590435

This work was approved and declassified by the Chemicals and Biotechnology Committee on 22/06/2026.

Please cite this publication as:

OECD (2026), *Validation report for Test Guideline No. 256 on determination of effects on earthworms (Oligochaeta, Annelida) in field studies,* OECD Series on Testing and Assessment, No. 430, OECD Environment, Health and Safety, Paris, https://one.oecd.org/document/ENV/CBC/MONO(2026)8/en/pdf.

Contact us

OECD Environment Directorate,

Environment, Health and Safety Division

2 rue André-Pascal

75775 Paris Cedex 16

France

E-mail: ehscont@oecd.org

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 38 countries in North and South America, Europe and the Asia and Pacific region, as well as the European Union, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several Partner countries and from interested international organisations attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in twelve different series: **Testing and Assessment**; **Good Laboratory Practice and Compliance Monitoring**; **Pesticides**; **Biocides**; **Risk Management**; **Harmonisation of Regulatory Oversight in Biotechnology**; **Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents; Safety of Manufactured Nanomaterials;** and **Adverse Outcome Pathways.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (https://www.oecd.org/en/topics/chemical-safety-and-biosafety.html).

[Para style] This publication was developed in the IOMC context. The contents do not necessarily reflect the views or stated policies of individual IOMC Participating Organizations.

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The Participating Organisations are FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank, Basel, Rotterdam and Stockholm Conventions and OECD. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

Foreword

This document describes the results of the validation effort for the method included in the new Test Guideline (TG) for the determination of effects on earthworms (Oligochaeta, Annelida) in field studies. This TG is designed to be used for assessing short- and long-term effects of chemicals on earthworms in soils under field conditions and is intended to be used for the registration of chemicals, probably most prominently for the testing of pesticides in the European Union. The TG is based on an existing ISO Guideline, i.e. ISO 11268-3, which has been in use since 1999.

The new TG and this accompanying validation report are the result of Test Guidelines Programme (TGP) project 2.47, which was included in the TGP workplan in 2013, and led by Germany. A dedicated Expert Group supported the development of these documents on several occasions since 2013, the latest in May 2024.

The subsequent Working Party of National Coordinators of the Test Guidelines Programme (WNT) review took place from November-December 2025, and the WNT approved the TG and this validation report at its 38th meeting in April 2026.

Table of contents

FIGURES

TABLES

Background and Intention

This document has been written in support of the submission of a new OECD Test Guideline, focusing on the study of the effects of chemicals (in particular plant protection products (PPPs)) on earthworms in the field as required by, e.g. European, authorities for the authorisation of chemicals (EC 2001; EC 2009). Since 1994, the risk of chemicals for earthworms in the field has been assessed by a test that had originally been standardised by the German Federal Biological Institute (No. VI, 2-3, BBA 1994). This document was transformed into an international guideline (ISO 11268-3) by the International Standardisation Organisation (ISO 1999). It has been updated several times up to now without changing the basic approach (ISO 2014). However, ISO guidelines usually focus on the assessment of (potentially) contaminated compartments (water bodies, sediments, waste materials as well as soils), i.e. they are used for retrospective environmental risk assessment approaches. In contrast, OECD guidelines serve the purpose of a prospective assessment of individual chemicals and defined chemical mixtures such as PPP formulations. As a consequence, ISO guidelines used in the testing of chemicals were transcribed to OECD format during the past 10 years.

In the course of this conversion, which was performed under German lead since 2013 as OECD project no. 2.47 (‘New Test Guideline on Determination of Effects on Earthworms in Field Studies’), it was checked whether apart from formal adjustments further changes were considered necessary. This assessment was performed by an ad hoc sub-group (led by Dr. J. Römbke) of the Global Soil Interest Group (GSIG) of the Society for Environmental Toxicology and Chemistry (SETAC). In this working group, experts from contract research organisations (CROs), industry, universities and public authorities have been working together for about seven years. Due to the experiences made in many regulatory tests as well as in some research studies during the past 20 years (e.g. Kula et al. 2006), the preparation of an improved earthworm field test guideline written according to the requirements of OECD was recommended. Besides methodological details, primarily the study design and the statistical evaluation of the test results had to be optimised. To address these issues, scientifically robust and practical information was missing. On behalf of the German Federal Environmental Agency (UBA) a project was launched in 2016 in order to close these knowledge gaps (in particular, test design and statistics, (Daniels et al. 2020; Römbke et al. 2020) and its results have been incorporated in this document.

In its structure and content, this document follows the requirements as laid down in the OECD Guidance Document No. 34 (2005). The validation process was performed in a retrospective way, i.e. no ring test[1] was performed because the earthworm field studies have been conducted for more than 20 years according to the existing ISO standard (ISO 1999; 2014) and the main procedures, processes of the study as well as the study set up have not changed. Moreover, field tests cannot be standardized to the

---

[1] A ring test is a method of external quality assurance for measurement procedures as well as measurement and testing laboratories. In principle, identical tests are conducted in different institutions or companies. Comparing the results allows for conclusions to be drawn, for example, about the measurement accuracy or measurement quality of the participating institutes or instruments, or about the quality of a product. Ring tests are, among other things, used to validate regulations for measurement procedures.

same extent as laboratory tests due to the natural variability of environmental conditions (e.g. weather conditions, soil properties, or the composition of the respective earthworm community) also meaning very high requirements in terms of resources (time and costs).

Therefore, validation activities described here are based on experiences gained in the practical performance and regulatory assessment of earthworm field studies, conducted in different European countries with several field situations in terms of soil characterisation, history of use, earthworm communities and tested chemicals.

In the following, OECD requirements as laid down in GD 34 (2005) were used as a yardstick for the preparation and validation of a new OECD Guideline on the use of earthworms for the testing of chemicals in the field. A guideline has been prepared in parallel to the activities described here.

Available dataset of field studies on earthworms

Existing data on field studies, conducted with the same reference substance Carbendazim or Benomyl cover different field situations in terms of location (country & climate), soil characterization (pH and C-org), history of tested sites (agricultural use type, formerly applied pesticides), community of earthworms as well as tested chemicals (different types of plant protection products: herbicides, fungicides and insecticides). In the course of the preliminary analyses and investigations, the ISIS database ("Information System Chemical Safety") of the German Environment Agency (UBA) was identified as useful source for data analysis of existing earthworm field tests, which held 150 entries for field studies on earthworms.

To demonstrate the large variability of habitats, test chemicals as well as community compositions, a subset of 70 studies (around half of all available studies at UBA) was analyzed to describe the variety of occurring environmental parameters as well as test chemicals.

The subset of 70 field studies on earthworms was conducted in six European countries: Germany, Denmark, United Kingdom, Sweden, Spain and France. The study sites have been in agricultural use before study start whereby the history included uses either as arable fields or as grassland sites before study start. Soil parameter varied in soil textures as well as in main soil parameters like pH (range: 4,7 – 7,9) and Organic carbon (range from 0,8 – 14,4 %). Also, biological parameters varied, described by the minimal number of earthworm species which ranged from 2 until 14, the total abundance ranged from 77 – 767 ind./m² and total biomass from 17 – 338 g / m² at study start. As reference substance Carbendazim or its main metabolite Benomyl was used, showing effects of at least 50 % on total abundance and / or biomass after application for rates of 4 – 10 kg Carbendazim/ha. Test chemicals differed strongly within the subset of 70 studies, only one test chemical was tested three times. All other 67 test chemicals were tested only once and were distributed within the three major groups of herbicides, fungicides and insecticides2.

[2] Due to company and business secrecy, more information cannot be reported within this document.

Table 1. Main parameters of a subset of 70 earthworm field studies. Study location in terms of country, land use before study start, sampling days, number of species at test start, abundance and biomass at test start as well as soil parameters (soil texture, soil pH and content of organic carbon. Studies shaded in grey have been used for detailed analysis within Römbke (2020).

No.	Sudy ID	Country	Land use before study start	No. Of sampling dates	No of species at test start	Mean individuals/m² at test start	Mean biomass g/m² at test start	Soil texture (USDA)	Soil pH	Organic Carbon [%]
1	test1000	UK	arable land	12	11	not possible	not possible	sandy loam	7.1	2.2
2	test1001	DE	grassland	4	7	129	63	high clay silt	6.2	1.7
3	test1002	DE	arable land	6	9	368	38	silt	6.7	2.8
4	test1003	UK	grassland	4	3	203	338	clay loam	6.7	4.9
5	test1004	DE	grassland	4	11	123	23	silty loamy sand	6.6	1.4
6	test1005	UK	arable land	4	13	192	99	sandy loam	6.3	1.6
7	test1006	DE	grassland	4	6	180	14	sandy loam	6.7	1.3
8	test1007	DE	arable land	6	5	n.d.	n.d.	sandy loam	6.5	1.5
9	test1008	DE	arable land	10	6	142	33	sandy loam, silt	n.d.	n.d.
10	test1009	DE	grassland	4	2	< 100	n.d.	n.d.	n.d.	n.d.
11	test1010	DE	arable land	4	5	195	51	loamy sand	7.9	n.d.
12	test1011	DK	grassland	5	6	99	83	not mentioned	n.d.	n.d.
14	test1012	UK	arable land	11	3	n.d.	n.d.	loamy sand	n.d.	n.d.
15	test1013	DK	grassland	4	7	120	82	loamy sand	5.7	1.0
16	test1014	DE	grassland	4	7	100	32	silty sand	5.8	2.0
17	test1015	DE	grassland	4	7	100	32	silty sand	5.8	2.0
18	test1016	DE	grassland	4	7	100	32	silty sand	5.8	2.0
19	test1017	DE	grassland	3	7	n.d.	n.d.	silty sand	5.7	1.8
20	test1018	DE	grassland	4	4	334	202	silty loamy sand	5.6	1.6
21	test1019	DE	grassland	4	5	196	29	loamy sand	6.4	1.6
22	test1020	DE	arable land	6	8	223	18	loamy silt	7.2	2.3
23	test1021	DE	grassland	15	14	296	17	Sandy silt & Clay loam	6.4	3.4
24	test1022	UK	grassland	4	9	263	298	silt loam	6.2	4.2
25	test1023	DE	arable land	4	4	77	17	loamy sand	5.4	0.8

26	test1024	DE	grassland	4	7	120	79	sandy, loamy silt	7.0	1.4
27	test1025	UK	grassland	2	11	172	102	sandy clay loam	7.1	2.2
28	test1026	DE	grassland	3	6	97	14	silty sand	6.5	2.4
29	test1027	UK	grassland	5	9	214	n.d.	Clay loam	5.4	4.3
30	test1028	DE	arable land	1	5	196	52	loamy sand	7.9	3.2
31	test1029	DE	grassland	4	9	386	237	medium loam sand	5.6	2.0
32	test1030	DE	grassland	4	8	154	83	loamy sand	6.0	1.6
33	test1031	DE	grassland	4	6	220	26	loamy sand	6.7	3.7
34	test1032	DE	arable land	3	5	328	98	silty loamy sand	6.0	1.2
35	test1033	DE	grassland	4	8	449	314	silt loam	5.4	5.3
37	test1034	DE	grassland	6	9	269	82	silty loam	7.2	2.1
38	test1035	DE	grassland	6	11	228	201	sandy silt	7.1	1.9
39	test1036	DE	arable land	4	7	312	124	pure silt/pure clay silt	6.9	2.1
40	test1037	DE	grassland	6	6	93	52	silty clay	5.1	1.1
42	test1038	DE	grassland	4	7	258	69	poor sandy loam	6.7	1.0
43	test1039	DE	grassland	4	6	103	34	loamy sand	5.8	1.3
44	test3064	DE	arable land	4	2	108	n.d.	clayey silt	7.0	1.2
45	test1041	DE	grassland	4	8	191	55	silty loam	5.4	2.2
46	test1042	DE	grassland	4	7	191	55	silty loam	5.4	2.2
47	test1043	DE	arable land	4	6	160	65	silty loam	5.4	1.0
48	test1044	DE	grassland	6	8	478	158	clay	6.9	3.9
49	test1045	FR	arable land	4	12	217	104	silty sand	6.3	0.9
50	test1046	UK	grassland	5	8	186	69	Clay loam	6.2	4.4
51	test1047	ESP	arable land	6	8	481	98	Sandy silt	5.1	3.9
52	test1048	ESP	arable land	8	4	206	47	Loamy silt	6.4	1.5
53	test1049	ESP	arable land	6	9	767	165	Silty sand	4.7	4.7
54	test1050	DE	arable land	9	10	305	99	silt	6.2	1.4
55	test1051	SE	arable land	5	9	106	46	loamy silt	6.4	1.8
56	test1052	UK	arable land	6	8	440	42	Silty clay	5.9	3.3
57	test1053	DK	arable land	5	7	116	83	loamy sand	6.8	1.3
58	test1777	DE	arable land	3	7	105	78	n.d.	6.5	1.7
59	test1941	DE	arable land	3	6	131	70	sand	5.4	1.3

60	test2225A	DE	arable land	6	11	368	177	sandy loam	6.7	1.6
61	test2225B	DE	arable land	6	12	368	177	sandy loam	6.7	1.6
62	test2225C	DE	arable land	6	11	368	177	sandy loam	6.7	1.6
63	test2237	DE	arable land	4	12	194	99	sand	6.3	n.d.
64	test2515	DE	grassland	4	13	344	331	loamy sand	6.9	3.9
65	test2594	DE	arable land	3	14	307	118	sand	5.1	1.3
66	test2674	DE	bare soil	4	10	202	149	Sandy loam	7.7	2.6
67	test2740	DE	arable land	4	14	307	118	n.d.	5.1	1.3
68	test2764	DE	bare soil	4	10	191	157	sandy loam	7.7	14.4
69	test2863	DE	bare soil	3	9	428	161	sand	6.6	1.8
70	test2921	DE	grassland	4	13	442	108	loamy sand	7.4	2.4
	Avarage			**4.8**	**8.0**	**236**	**103**		**6.3**	**2.4**
	Minimum			**1.0**	**2.0**	**77**	**17**		**4.7**	**0.8**
	Maximum			**15.0**	**14.0**	**767**	**338**		**7.9**	**14.4**

Note: n.d. = not detected. Grey shaded 21 studies have been used for detailed statistical analyses within Römbke et al., 2020. For further information see chapter 3.2.

Validation of the existing test method and data

For a detailed statistical analysis within the project of Römbke et al. (2020), quality criteria for data were defined with regard to further statistical investigations.

For the improvement of the existing design, only earthworm field studies possessing the following characteristics were used for detailed statistical analyses:

1. Earthworms should have been sampled by a combination of formalin/allyl isothiocyanate (AITC) extraction and hand-sorting. A bias of the sampled species composition and total abundance due to the use of the octet sampling was therefore prevented.
2. 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.
3. The field studies were carried out according to ISO 11268-3 or in line with BBA guideline part VI, 2-3.

21 field studies fulfilled these severe characteristics (shaded grey in **Error! Reference source not found.**). T o enlarge the database, further randomly chosen 5 studies with digitalized raw data at sample or plot level were integrated into the database, each with a slightly different sampling method. This data, however, was not used for meta-analyses of the species composition, but only for feasible test-internal statistical evaluations. 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.

Species composition

The sampled earthworm individuals of the 21 field studies belong to a total of 17 different species (**Error! R eference source not found.**). In addition to these analyses, aggregated taxa groups and genus level juveniles were subdivided. Aggregated groups were developed for “Total earthworms”, “Total adults”, “Total juveniles”, “Total anecics”, “Total endogeics”, “Total epigeics”, “Total epilobous adults”, “Total epilobous juveniles”, “Total tanylobous adults” and “Total tanylobous juveniles”. Undetermined individuals were excluded from further analyses. A detailed calculation of the natural variability of earthworm biomass and abundances, illustrated by the coefficient of variation for control treatments, is calculated and presented in comparison to results of the experimental field study in Römbke et al. (2020).

Table 2. Percentages of sampled earthworm species and assigned ecological and morphological groups

Species	dominance [%]	ecol. group	morph. group
Aporrectodea caliginosa (Savigny, 1826)	46.3	endogeic	epilobous
Aporrectodea rosea (Savigny, 1826)	17.1	endogeic	epilobous
Lumbricus terrestris Linnaeus, 1758	11.9	anecic	tanylobous
Allolobophora chlorotica (Savigny, 1826)	9.6	endogeic	epilobous
Lumbricus castaneus (Savigny, 1826)	4.4	epigeic	tanylobous
Aporrectodea longa (Ude, 1885)	2.9	anecic	epilobous
Lumbricus rubellus Hoffmeister, 1843	2.5	epigeic	tanylobous
Aporrectodea limicola (Michaelsen, 1890)	2.1	endogeic	epilobous
Octolasion tyrtaeum (Savigny, 1826)	1.7	endogeic	epilobous
Murchieona minuscula (Rosa, 1906)	0.883	endogeic	epilobous
Octolasion cyaneum (Savigny, 1826)	0.5	endogeic	epilobous
Proctodrilus antipae (Michaelsen, 1891)	0.052	endogeic	epilobous
Eisenia fetida (Savigny, 1826)	0.027	epigeic	epilobous
Satchellius mammalis (Savigny, 1826)	0.019	epigeic	epilobous
Aporrectodea cupulifera (Tetry 1937)	0.018	endogeic	epilobous
Lumbricus festivus (Savigny, 1826)	0.005	endogeic	tanylobous
Dendrobaena illyrica (Cognetti, 1906)	0.003	epigeic	epilobous

According to Römbke (2020), descriptive metadata of the 21 field studies reveal that the composition of species among all field studies report 6 - 14 species per study. The respective Shannon Diversity Index was between 0.3 and 1.6 (mean: 1.2). The diversity index is 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 is at least 10. This trend can also be observed for the mean individuals sampled, which is about 372 ind./m² on grassland, 356 ind./m² on bare soil and about 196 ind./m² on crop sites.

Statistical power on species level as well as for aggregated groups

As a statistical measure, the minimum detectable difference (% MDD) for the endpoint total earthworm abundance between control and treatment of all field studies was calculated according to Brock et al. (2015), amended by Duquesne et al. (2020) (Figure 1 and 2). The MDD approach (Brock et al., 2015 & Duquesne et al. 2020) was used with regard to the common statistical evaluation of the investigated field study reports using hypothesis tests of the t-test family. Nevertheless, in some cases non-parametric methods were applied.

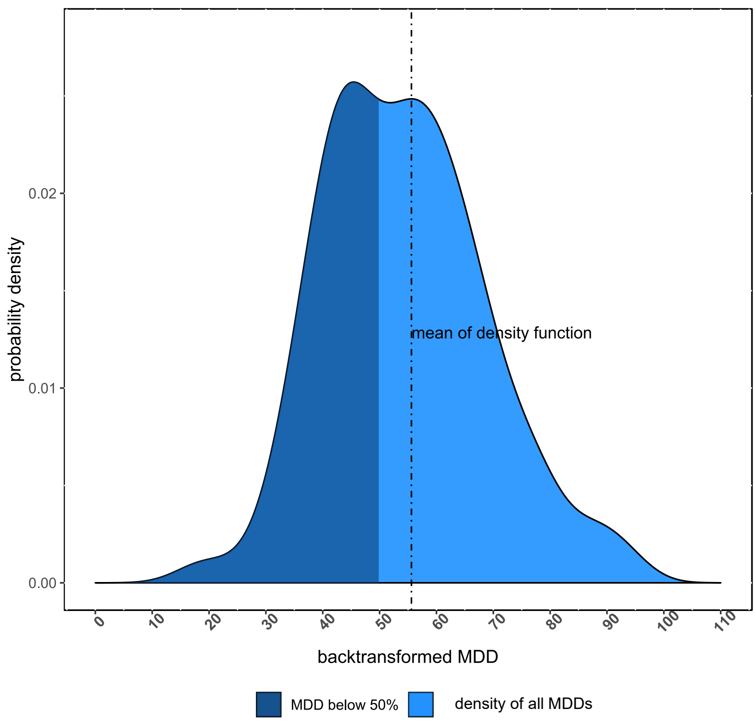
According to Römbke et al. (2020), the most likely value of the MDD for abundance data of total earthworms in the database is 45%, but the probability of obtaining a minimum detectable difference smaller than 50 % of the control is 42 %. The probability of obtaining an MDD between 10 % and 35 % was 8 % (small effects according to EFSA PPR (2017)). However, this range relates to effects on the protection goals and not to the measurement endpoints as the analysed earthworm field data. The same calculations for total biomass give even lower power values than for total abundance: a minimum detectable difference smaller than 50% was only detected in 32 % of the cases (mode at 67 %, which is the most likely value of the density function). Therefore, historical studies routinely lack adequate power and regulatory decision making is hampered by low statistical power.

Table 3. MDD values for total earthworms, abundance and biomass (if available), aggregated over all sampling dates for studies derived from the database.

	Abundance	Biomass
Most likely MDD	45 %	n.c.
Probability of detection for MDD < 50 %	42 %	32 %
Probability of detection for MDD > 10 % and < 35 %	8 %	n.c.

n.c. not calculated

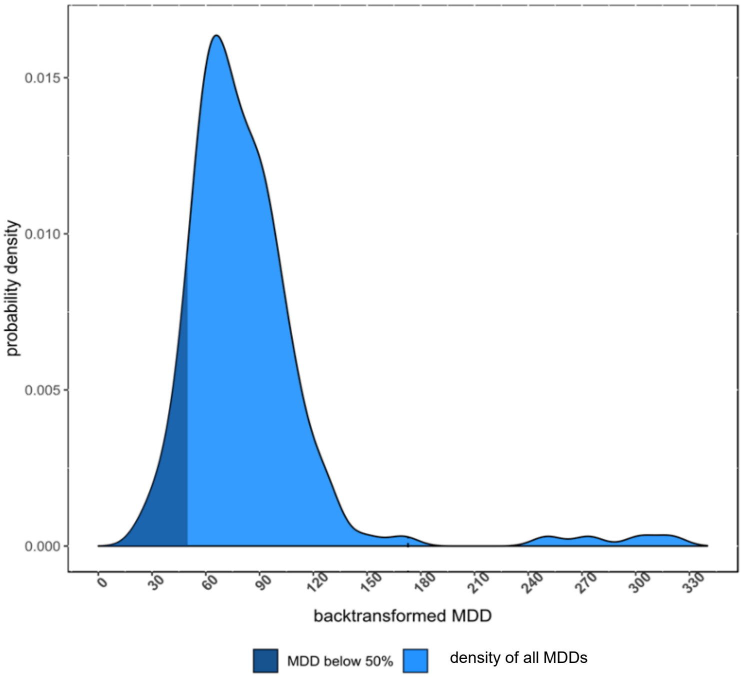
Figure 1. Distribution of the probability density for the minimum detectable difference (MDD in %) of total earthworm abundance data in the earthworm field study database extracted from ISIS (all sampling dates, empirical MDD between 11 % and 100.2 %)



Source: RWTH Aachen University

For the aggregated group of total earthworms, the most powerful MDDs have been calculated. Even for the most dominant species in the database, *Aporrectodea caliginosa* (46% of all sampled individuals), considerably lower probability was found to detect statistically significant effects in field studies (Figure 2).

Figure 2. Distribution of the probability density for the minimum detectable difference (MDD in %) of *Aporrectodea caliginosa* abundance data in the earthworm field study database (all sampling dates, lowest empirical MDD at 15.4 %)



Source: RWTH Aachen University

The probability to calculate a minimum detectable difference less than 50 % between treatment and control for the species *Aporrectodea caliginosa* is low, only 12 % of the probability density distribution can be assigned to this range. The most likely value of the calculated probability distribution for minimum detectable differences for abundance data of *Aporrectodea caliginosa* is 66 % (mode density function). This corresponds to a reduction of earthworm abundances of two third in treatments compared to a control. A less substantial reduction would not provide a statistically significant effect, even if biologically relevant according to EFSA (2019). Again, even lower minimum detectable differences were calculated for the endpoint biomass. An MDD smaller than 50 % was only calculated for 8 % of the data. The respective mode of the density function is about 70 %. Statistically significant identification of medium-sized effects is rare for this endpoint due to the variability of data. Therefore, regulatory decision making based on low statistical power is hampered also for the most dominant species.

Following table gives an overview for MDD values observed at the pre-sampling in the database studies for abundance and biomass. MDD values are presented for aggregated groups as well as the most frequent species.

Table 4. MDD values at pre-sampling for abundance and biomass for aggregated groups as well as most dominat species presented as mean and median including standard deviation.

	Minimum detectable difference - backtransformed data [%]					
endpoint measure	Abundance data - data base			biomass data - data base		
sampling point	pre- application			pre- application		
MDD backtransformed_orig	mean	median	st.dev	mean	median	st.dev
Allolobophora.chlorotica.ad.	132,19	150,00	52,06	160,03	180,60	73,68
Aporrectodea.caliginosa.ad.	59,84	53,80	29,37	64,50	56,20	32,38
Aporrectodea.rosea.ad.	83,29	76,20	31,83	92,61	86,10	42,98
Lumbricus.castaneus.ad.	123,64	115,20	41,44	153,41	142,35	54,80
Lumbricus.terrestris.ad.	115,30	75,30	91,24	102,79	70,55	70,21
total.adults	46,19	45,50	15,82	54,68	51,30	18,41
total.anecic.ad.	102,15	67,70	90,30	95,07	68,15	67,00
total.earthworms	44,22	41,35	17,20	48,15	47,80	17,38
total.endogeic	51,15	48,95	18,60	54,93	52,65	19,33
total.epigeic	115,43	112,35	43,83	134,90	126,25	56,83
total.epilobous.adults	50,07	48,40	17,81	57,85	54,20	24,20
total.epilobous.juveniles	56,90	51,10	24,13	58,54	50,00	31,64
total.juveniles	51,41	47,30	21,78	58,34	56,90	22,44
total.tanylobous.adults	72,36	66,60	25,21	75,70	67,95	27,39
total.tanylobous.juveniles	71,10	62,40	31,11	80,90	68,70	32,16

In an overall picture, best-practice studies (hand-sorting and formalin/AITC extraction) reveal low power to detect differences between control and treatment plots for aggregated taxa. For single species, this statistical potential for a reliable statistical identification of effects is even lower, as the example of *Aporrectodea caliginosa* shows. The even lower abundances of other species and the resulting increasing variability in the datasets (see Römbke et al. (2020) for details) lead to an even weaker distribution of the MDD for other single species. There is a chance of obtaining sufficient MDD for single taxa to identify small effects. However, the overall MDD for species and earthworm groups show, that a comprehensive detection of small effects is not given in analysed earthworm field tests, conducted according to ISO 11628-3 (2014). For all sampled species, an MDD of 35 % (threshold value for small effects) lies below the interquartile range in the distribution of the MDD considering all tests in the database. Across all sampling points, endpoints, species and earthworm groups of all database studies, only 8.4 % of the measured MDD are below 35 %. Therefore, an amended test design should contain an option to perform regression approaches as an alternative to the No observed effect concentration/No observed effect rate (NOEC/NOER) approach as well as an amended NOEC/NOER design.

Experimental check of study amendments - Performance of a field study and combined evaluation with data base studies

In a joint discussion between the UBA and the project consortium in respect to the results of the evaluation described above, the “mixed omni-design” design was proposed, characterized by combining a so-called NOEC/NOER- with an effect concentration x / effect rate x (ECx/ERx)-design. The “mixed omni-design” is characterized by:

- 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).

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.

Statistical analysis: single species versus aggregated groups

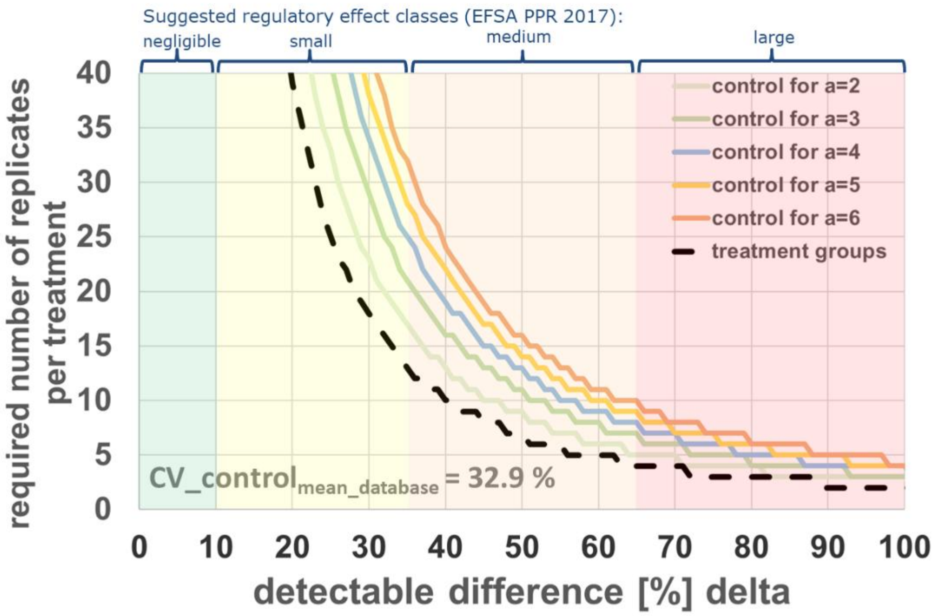
The main focus was to improve the conventional statistical methods to evaluate earthworm field studies (ISO 11268-3) and to acquire insights for statistical considerations regarding an adapted test design for earthworm field studies. Raw data for biomass and abundance at all sampling dates and for all 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 of the data from the experimental field study 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 chemicals based on an aggregated endpoint such as total abundance of all earthworms.

Required replicates

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 studies as well as the experimental field study 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 endpoints (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 chemical (= high MDD).

Römbke et al. (2020) used the theoretical sample size calculation for earthworm field studies according to Dunnett's multiple test procedure (Horn & Vollandt 1995). As an example of application, following Figure 3 illustrates the simulation of required replicate numbers that correspond to the mean coefficients of variation of the group "total earthworms" (= 32.9 %, field tests available in database). Test power was set to 80 %. The required number of replicates is plotted against the desired detectable difference between control and treatments with varying number of treatment levels.

Figure 3. Number of required replicates (plots) per treatment in earthworm field tests plotted against the detectable difference (in percent) between treatment and control. Variation of control was set to 32.9 %, which is the mean variability of total earthworms in available field studies. Coloured lines: Required replicates for controls using different numbers of test treatments (a); dotted line: Required replicates of test treatments



Source: RWTH Aachen University, extracted from Römbke et al (2020).

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. 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 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/NOER calculation using single samples (= subplots) as statistical replicates would result in an improvement with regard to MDDs. To identify the minimum sample size, the measured CVs on plot level (current standard method), and on sample level was used to assess shifts in test power. Results based on data of the field study showed that the increase in plot numbers (n=6) and the larger plot size (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 effects classified as ‘small (10 % - 35 %)’ by EFSA (2017) 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.

However, it has been discussed within experts in commenting rounds, that evaluation on sub-plot level could be classified as 'pseudo-replication'. Therefore, evaluation on sub-plot level is not recommended and evaluation on sup-plot level is only used for the identification of outliers.

Methods used for statistical evaluation

The calculation of the effect thresholds was carried out for all suitable database studies identified by Römbke et al. (2020) for all sampling times by comparing the Dunnett method with results from the so-called CPCAT approach (Closure Principle Computational Approach test). 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 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 chemical 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/NOER 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).

Dietrich and Wang (2025) performed power calculations for CPCAT, CPFISH, and Dunnett-GLM using simulated effect sizes to reflect typical reproduction and field test designs.

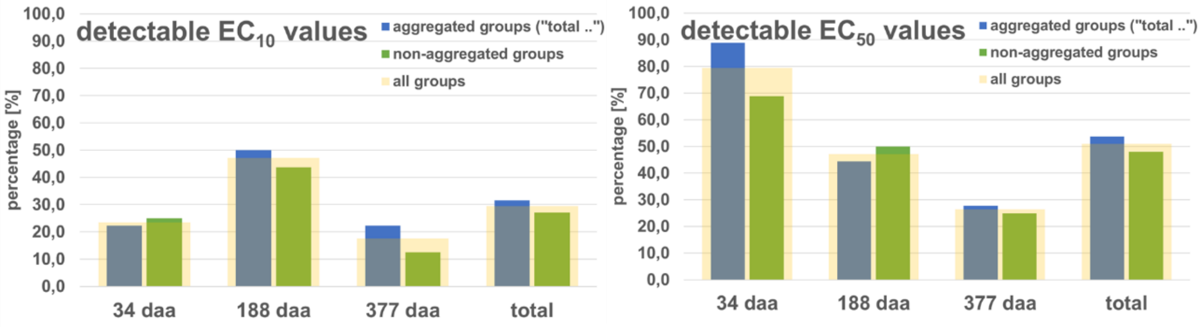
Across most tested scenarios, CPCAT and Dunnett-GLM exhibited only minor differences in statistical power and overall performance. Variations in power were observed for both methods when multiple treatments were included. In cases where substantial deviations from the Poisson assumption are detected, the application of a generalized linear model with Dunnett contrasts (GLM-Dunnett), assuming a quasi-Poisson distribution to account for increased dispersion, is recommended (Hothorn & Kluxen, 2020). This approach explicitly accommodates extra-Poisson variability and has been shown to provide consistent and robust power properties in scenarios of high variability compared to CPCAT (Dietrich & Wang, 2025).

Test design: Dose response versus NOEC/NOER

In contrast to the database studies, the experimental field 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/NOER approach (Dunnett test), the choice of an ECx/ERx 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 ECx/ERx 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.

Following figure 4 shows a comparison of the percentage of detectable EC10 and EC50 values during the test period:

Figure 4. Percentage of detectable EC10 (left) and EC50 values (right) of all calculated dose-response curves for endpoints of the earthworm pilot field study data. Daa = days after application.



Source: RWTH Aachen University, extracted from Römbke et al (2020).

Further information is captured in chapter 3.2.3 of Römbke et al (2020).

Community analyses

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 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.

Consequences for the study design and evaluation

The following basic recommendations and requirements regarding the implementation and evaluation of earthworm field tests were identified, based on the detailed analyses of data base field studies in combination with the experimental field study:

- Determination and evaluation of the parameters biomass and abundance should be done at species level, as the aggregated morphological or functional groups used may disguise effects on single species.
- The number of replicates needs to be raised due to high variability in the field. Six replicates should be used for control and reference plots, whereby the reference plots can be lowered to four plots in case of high earthworm abundances in the test field. Depending on the test design, replicates for treatments vary between two and six plots (see table 3).
- The ECx/ERx design is a meaningful alternative to the NOEC/NOER design in the earthworm field test. At least a mix design would be advisable. In fact, the ECx/ERx 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/NOER evaluation is avoided.
- The calculation of effect thresholds like NOEC/NOER or lowest observed effect concentration / lowest observed effect rate (LOEC/LOER) 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 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.

- The application of a generalized linear model with Dunnett contrasts (GLM-Dunnett), assuming a quasi-Poisson distribution to account for increased dispersion, is recommended for non-equal dispersed data.
- 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 (ECX/ERX) 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.
- Principal response curves are generally applicable within the ECX/ERX-design and a powerful tool for community analyses. They should be carried out in addition to univariate methods when appropriate data are available, for tests with multiple treatments (e.g. ECX/ERX design).
- A flow chart for the choice of the statistical method is included in the test guideline in section 'analysis of data / evaluation of results'.

Proposed test method protocol

The experience gained during the project of Römbke et al. (2020) in combination with two experts' meetings as well as two commenting rounds within the scientific community were applied to derive a proposal for an amended test designs in the following table.

Table 5. Number of plots and treatments for the ECx/ERx-, the mixed and NOEC/NOER/NOER design in earthworm field tests. C control; T 1-x treatments; R reference substance

Test design	Plots per test treatment (No.)									Plots (sum)	Samples (total No.)
	C	T1	T2	T3	T4	T5	T6	(T7)	R		
EC/ERx Design	6	3	3	3	3	3	3	(3)	6	30 (33)	120 (132)
Mixed Design	6	2	6	2	2	6	-	-	6	30	120
NOEC/NOER Design	6	6	6	6	-	-	-	-	6	30	120

Proof of the eight OECD validation criteria

According to the OECD Guidance document on the validation and international acceptance of new or updated test methods for hazard assessment (2005, GD 34) the listed 8 criteria are discussed as follows:

Criterion 1: Rationale for the test

The performance of an earthworm field test is scientifically suitable when being interested in the effects of chemicals, in particular pesticides, on the soil invertebrate community. The choice of earthworms as representatives of the soil invertebrate community is justified due to their high ecological value in terms of the ecosystems services they provide (Mulder et al. 2011; Brussaard 2012).

In Europe the earthworm field test is triggered in case that the intended use of a chemical (either an active ingredient or a formulation) poses a risk to earthworms, as determined in a laboratory test observing the earthworm reproduction according to OECD 222 (2004). For plant protection products (PPP), the requirement of an earthworm field test is captured in the corresponding data requirements (EC 2013 a & b) as well as in associated guidance documents on Terrestrial Ecotoxicology (EC 2002).

Currently, the Earthworm Field Test is the only available standard test method for studying the effect of pesticides (or other chemicals) on soil invertebrates in the field (EC 2009; EC 2013). Therefore, the results of such a study are of utmost importance regarding the question whether the pesticide being tested may pose a risk for the environment.

Criterion 2: Relationship of the endpoint(s) determined by the test method to the biological effect

The overall aim of the earthworm field test is to determine whether the application of the tested chemical adversely affects the respective total earthworm community, single species, ecological or morphological groups at the experimental site. The most appropriate endpoints to be determined in this test are the abundance, biomass and diversity of earthworms. Effects on these endpoints also indicate whether the functional role of the earthworm community (i.e. their contribution to the provision of soil functions and services) after an application of a chemical has been affected (EFSA 2017). Thus, there is a direct relationship between the measured endpoints in this test and the biological effect data to be used for environmental risk assessment for this specific group of organisms. However, also other soil organisms with important contribution to soil function and services might be affected by the application of chemicals on soils, too. The effects on other in-soil organisms cannot be described on the base of an earthworm field study.

Criterion 3: A formal detailed protocol must be provided and should be readily available in the public domain

A formal detailed protocol is available since a first guideline was published back in 1994 (BBA (Biologische Bundesanstalt), 1994). Since then, the guideline has been transferred to the international level (as ISO 11263-3, 1999) and updated after intermediate expert meetings (described in Kula et al. 2006). A revised standard version is now available as ISO 11263-3 (2014). The presented current OECD Test Guideline is based on the most recent ISO version. In the framework of a research project (German Environment Agency, UBA FKZ 3715 67 420 0 (Römbke et al. 2020), the existing test design was improved and the statistical data evaluation was reviewed and improved based on the available data collected since the 1990ies. In addition, also the technical performance of the test was modified and improved. Due to the inclusion of many experts from government, industry and academia in the planning and implementation process during the project (e.g. during meetings in Nantes (France) 2016, in Flörsheim (Germany) 2017, and in Dessau (Germany) 2019), a broad input of expertise was secured. All data are available on request from the German Federal Environmental Agency (Dessau), referring to the Project No. (FKZ) 3715 67 420 0. Details on the practical performance of the pilot study as well as the assumptions and calculations used in data analysis as well as in the statistical analysis are laid down in the final report of this project (Römbke et al. 2020).

Criterion 4: Intra-test variability, repeatability and reproducibility of the test method within and amongst laboratories should have been demonstrated

Data on the performance of earthworm field studies in the open literature addressing variability, repeatability and reproducibility is available to a limited extent, only. One characteristic element of field studies is that the specific situation in which a test is performed cannot be completely reproduced at another site. A dedicated inter-laboratory comparison study (i.e. ring test) has not been performed. Still earthworm field studies according to the ISO standard have been performed for the last 20 years and more than 150 studies have been submitted for PPP assessment in Germany alone. Therefore, the study was successfully performed on at a broad range of field sites with diverse conditions (country, climate, soil conditions like soil type, pH, and Organic carbon content as well as earthworm population, cultivation, etc.) as shown in chapter 2. The reproducibility and repeatability of earthworm field studies is demonstrated by the toxic reference (products containing the active substances carbendazim or benomyl), as it was run in every performed test.

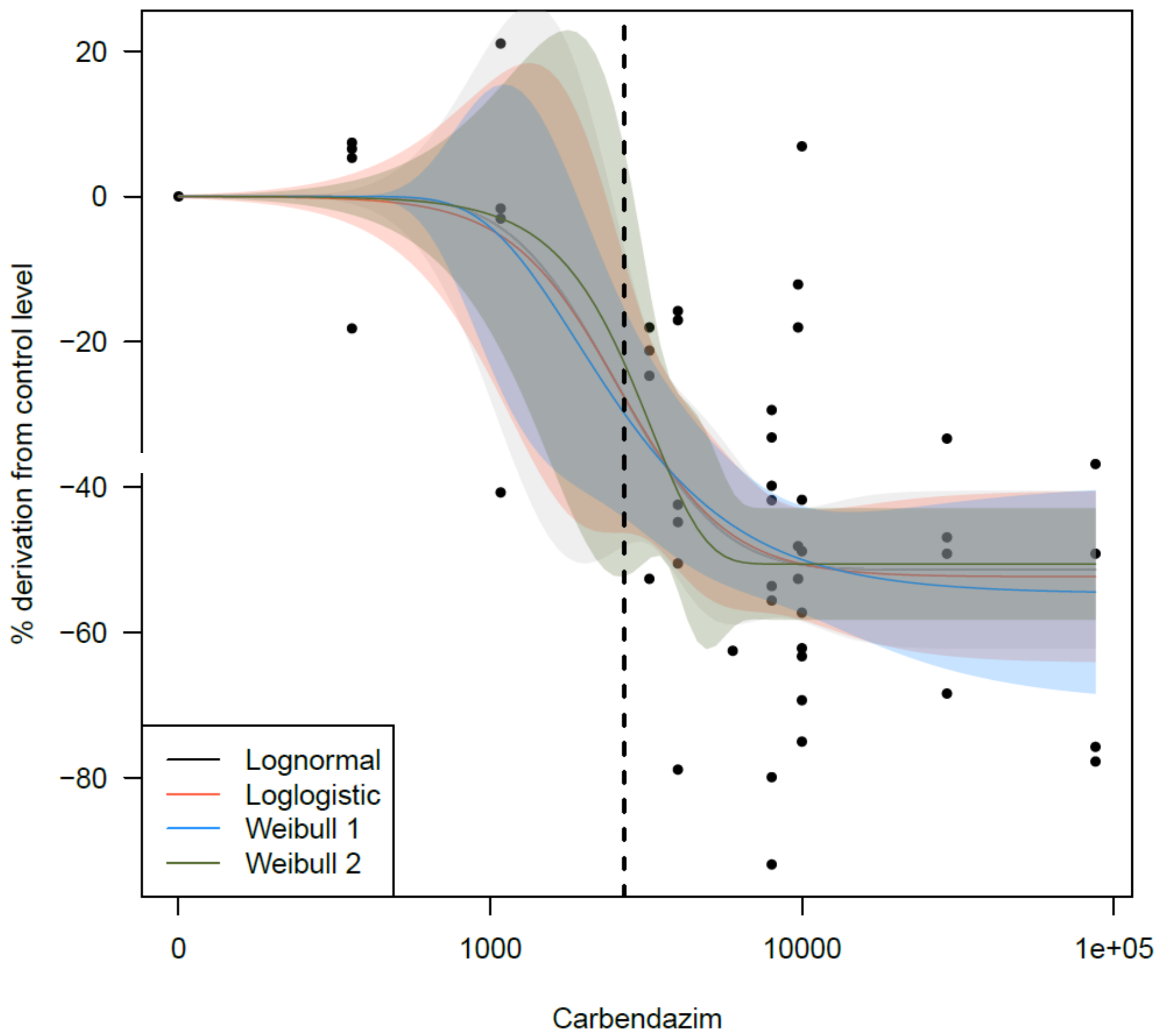
Therefore, in the following, three approaches are summarized:

- Statistical evaluation and literature example on the effects of carbendazim on earthworms in the field (Chapter 0);
- Critical evaluation of the methodology (Chapter 0);
- Results and recommendations for amendments of the current design (Chapter 0).

Examples on the effects of carbendazim on earthworms in the field

The effect of Carbendazim has been checked in nearly all available field studies within Europe as it is the recommended reference substance for earthworm field studies according to ISO (2014, see chapter 0). Figure 5 shows the deviation from the control in % on total abundance of earthworms 4 – 6 months after application of different rates of carbendazim of a data subset of 21 earthworm field studies as well as one study from open literature (Römbke et al. 2004):

Figure 5. Effects of carbendazim within a subset of earthworm field studies. Dose-Response has been tested with 4 modells: Lognormal, Loglogistic, Weilbull 1 and Weilbull 2



The example above shows an existing dose-response on the parameter total abundance of earthworms due to the application of Carbendazim. The variation in response is caused by natural variability as well as shortcomings within the current test design, described in chapter 0 of this document.

Critical evaluation of the methodology

General comments

(1) Most of the earthworm field tests so far have been performed in central Europe (i.e. the Atlantic and Continental biogeographical regions), few in the Mediterranean, almost none in the boreal region and none at all in Eastern Europe (Dinter et al. 2013). Thus, the conclusions drawn are valid for this region and its respective earthworm community, but not for other regions of the world (especially the tropics but also Northern regions) with their completely different earthworm communities. As far as we know no detailed

regulatory requirements regarding earthworm testing in the field exist outside of Europe, meaning that information on such effects is limited (Foerster et al. 2006).

(2) Taxonomic information (including identification keys) for earthworm species is well available, e.g. Bouché (1972), Sims & Gerard (1999) and Sherlock (2018) for Western and Northern Europe (especially the British Isles), Krück (2018) for Germany and Csuzdi & Zicsi (2003) for South-Eastern Europe. However, outside of these areas the identification of earthworm species may be difficult due to a combination of a high number of endemic species (i.e. species which occur only regionally), which are often not well-known ecologically and/or even not yet scientifically described (Lavelle et al. 1999).

(3) In the past, the identification of earthworm species depended on the availability of a limited number of taxonomic experts, which limits the routine application of this method. However, the use of newly developed DNA-based genetic methods for species identification is evolving rapidly – both for individual species as well as for eDNA extracted from soil samples (e.g. Pérez-Losada et al. 2009, James et al. 2010, Richard et al. 2010, Jänsch et al. 2025; Cuartero et al. 2025). These methods may become a relevant alternative within the next decade.

(4) When selecting test sites areas with “extreme” soil properties, e.g. low pH-values (Dinter et al. 2012) should be avoided, since at such sites abundance and diversity of earthworms is probably very low, meaning that robust test results are difficult to obtain.

(5) One characteristic element of field studies is that the specific situation in which a test is performed cannot be completely reproduced at another site and date due to differences in soil properties (e.g. including their spatial distribution in the field scale or below) or climatic conditions (and thus the size and/or composition of the earthworm community).

Finally, agricultural activities will differ to a certain extent, which may also influence the earthworm community. However, experience in the last 20 years has shown that despite these limitations the basic question (i.e. does this chemical affect the earthworm community under field conditions) can be answered in a robust way only by performing a standardized earthworm field test.

Analysis of a subsample to investigate statistical issues

Until now, a broad range of pesticides has been tested with earthworm field studies. Within the analysed subset of data (see chapter 3), 14 herbicides, 10 fungicides, and 5 insecticides have been evaluated within Römbke et al. (2020). The analysation of the database studies as well as the evaluation of the experimental field studies shows high shortcomings of the current test design due to low statistical power of earthworm field studies (see Chapter 3).

Suggested amendments for earthworm field studies

All details of the evaluation of a data subset in combination with an experimental study are described in (Römbke et al. 2020), a short overview of the new design options is given in chapter 0 of this document. The purpose of the 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 amended test design should include a raised number of replicates (up to six plots depending on the chosen design) and may contain an option to perform regression (ECx/ERx) approaches, which have been suggested as an alternative to the currently performed threshold (NOEC/NOER) approach. In general, the ECx/ERx design leads to more robust conclusions for environmental risk assessment. Alternatively, an amended NOEC/NOER design with 6 replicates each may be performed. The calculation of effect thresholds (NOEC/NOER/LOEC/LOER) should be conducted with the most powerful and appropriate multiple test procedure for given data prerequisites. If applicable to the data, the closure principle computational approach test (CPCAT) is the preferred option for count data. Parallel use of different statistical tests may help to interpret the data sets. According to the experiences made during the performance of the experimental study and the results of the statistical analyses, an OECD test guideline was developed.

Criterion 5: The test method’s performance must have been demonstrated using a series of reference chemicals preferably coded to exclude bias

This requirement is especially referring to the testing of chemicals with a known mode-of-action to be used in in-vitro methods. Thus, it is not relevant for earthworm field studies, since this is not an in-vitro method. In fact, to the best of our knowledge no data on the toxicity of a reference chemical on earthworms in the field have been generated so far, with the exception of the two – closely-related - reference chemicals benomyl and carbendazim. Almost all of these tests have been performed with just one concentration. In several parts of the world, carbendazim is not authorised as plant protection product anymore. However, it’s use as a reference chemical in earthworm field trials is still possible with special permission from regulatory bodies.

However, the performance of the earthworm field test has been thoroughly demonstrated in the past, since this test has been used in regulatory soil ecotoxicology for about 20 years, covering a wide range of test chemicals (mainly pesticides with different mode of actions, but also metals). As been discussed in detail (see, e.g., Daniels et al. 2020) there is a need for optimizing the test design as well as the statistical handling of the test results, but this test has regularly been performed:

- At various sites, with different soil properties and earthworm communities;
- With a wide range of test chemicals from chemical classes;
- By different institutions, such as governmental agencies, companies, contract laboratories (mostly under GLP) but also at universities (in a scientific context).

Criterion 6: The performance of test methods should have been evaluated in relation to existing relevant toxicity data as well as information from the relevant target species

This point does refer to the case in which a substitute test method is developed. In this case, adequate data should be available to permit a reliable analysis of the performance and comparability of the proposed substitute test method with the test it is designed to replace. The earthworm field test method that addresses the effect of chemicals on a community of organisms in the field is not designed to replace any already existing test. In fact, all other ecotoxicological test methods with earthworms are laboratory tests, which use one species (*Eisenia andrei*/*fetida*) which does not naturally occur in agricultural soils. In addition, the existing laboratory tests differ clearly from those performed in the field: for example, in the laboratory an artificial soil is used, not a natural soil as in the field. In addition, environmental conditions like temperature are kept constant at 20°C, while in the field temperature does vary, usually in a lower range than the 20° C. Thus, the earthworm field test is not a substitute test method.

Criterion 7: All data supporting the assessment of the validity of the test methods including the full data set collected in the validation study must be available for review

From the very first beginning of earthworm field studies, i.e. for more than 20 years, two systemic benzimidazole fungicides have been used as reference substances. Their high toxicity for earthworms became evident already in early seventies after their “normal” use in apple plantations (Stringer and Wright, 1973). About 10 years later Edwards and Brown (1982) proposed the first ideas regarding testing earthworms in the field. Thus, in the first BBA Guideline on the laboratory testing of earthworms benomyl or its first metabolite carbendazim has been included as a reference substance (BBA 1984) and in the subsequent ISO guidelines (ISO 1998; 2014). As far as we know no other reference chemical has ever been proposed, despite the fact that within the last 5 years such an alternative was searched for by soil ecotoxicologists, since due to legal restrictions the two existing ones are not available in the European Union anymore. In addition, the data gained in these reference tests are not useful in this context since they are almost always performed according to a limit test design, i.e. one concentration was tested against a (water) control. However, as already mentioned (Römbke et al. 2020) the effects of Carbendazim have been studied in an extended earthworm field test (= pilot study), i.e., a study which was performed according to ISO 23611-3 (2014).

All data collected in the Pilot Study are available on request from the German Federal Environmental Agency (Dessau), citing the Project No. (FKZ) 3715 67 420 0. They are summarized and discussed in the final report of the UBA project (Römbke et al. 2020). Parts of this work have also been published in the international scientific literature (e.g. Daniels et al. 2020).

Criterion 8: The data should have been obtained in accordance with OECD GLP Principles

In the European Union, requirements regarding OECD GLP principles are in practice for more than 25 years (e.g. EC 2009). At least since the first guideline (BBA 1994) on Earthworm Field Tests was published, data on the possible toxicity of chemical to earthworm in the field have been asked for by various regulatory authorities. To our knowledge, all data provided to governmental authorities in the context of pesticide registration were performed under GLP. However, it should be noted that the Pilot Study mentioned already in Chapter 3 followed as far as possible the respective guidelines but was not a GLP-study.

Summary and Outlook

Despite the fact that these eight criteria were originally not intended to be used for complex field studies it could be shown that they have been fulfilled – as far as possible for such a very specific test guideline. In addition, it has been accepted for more than 20 years by various European governmental agencies as decisive requirement for the registration of a broad range of pesticides. It is expected that the new version (improved mainly in terms of test design and statistical evaluation of the test results) will play this role for the decades to come.

References

BBA (Biologische Bundesanstalt für Land- und Forstwirtschaft) (1984): Verfahrensvorschlag: Toxizitätstest am Regenwurm *Eisenia fetida* in künstlichem Boden (LC 50, 14 Tage). Verfahrensvorschlag. (Nr. VI, 2-1), Braunschweig.

BBA (Biologische Bundesanstalt) für Land- und Forstwirtschaft (1994): Richtlinien für die amtliche Prüfung von Pflanzenschutzmitteln, Nr. VI, 2-3, Auswirkungen von Pflanzenschutzmitteln auf Regenwürmer im Freiland, Braunschweig.

Bouché, M.B. (1972): Lombriciens de France. Écologie et Systématique. – Paris (Institut National de la Recherche Agronomique): 671 pp.

Bouché, M.B. (1977): Stratégies lombriciennes. Ecol. Bull., 25: 122–132.

Brock, T.C., Hammers-Wirtz, M., Hommen, U., Preuss, T.G., Ratte, H.T., Roessink, I., Strauss, T. & Van den Brink, P.J. (2015): The minimum detectable difference (MDD) and the interpretation of treatment-related effects of pesticides in experimental ecosystems. Environ. Sci. Pollut. Res. Int. 22: 1160 – 1174.

Brussaard, L. (2012): Ecosystem Services Provided by the Soil Biota. In: Soil Ecology and Ecosystem Services (First Edition). Wall, D.H. et al. (Eds.). Oxford University Press, UK. Pp. 45-58.

Csuzdi, C. & Zicsi, A. (2003): Earthworms of Hungary (Annelida: Oligochaeta, Lumbricidae). Pedozoologica Hungarica 1: 1–272.

Cuartero, J., Briones, M.I.J., Rast, B.M., Stierli, B., Maurer-Troxler, C., Hug, A., Widmer, F., Schlaghamerský, J., Frey, B. (2025): Earthworm and enchytraeid indicator taxa of different land-use types identified using soil DNA metabarcoding, Applied Soil Ecology, Volume 206, 2025, 105891, ISSN 0929-1393, https://doi.org/10.1016/j.apsoil.2025.105891.Daniels, D., Jaensch, J., Roß-Nickoll, M., Roembke, J., Scholz-Starke, B. & Ottermanns, R. (2020): Application of the CPCAT to assess ecotoxicological field studies: Meta-analysis using earthworm field test abundance data. Envir. Toxico. Chem.

Darwin, C. (1881): The formation of vegetable mould through the action of worms with observations on their habits. Murray, London. 298 pp.

Dietrich, C. and Wang, M. (2025): Evaluation and implementation of statistical methods to assess effects in count data. Project No. 173522. Report No. (UBA-FB) FB001755/ENG. DOI: https://doi.org/10.60810/openumwelt-7762

Dinter, A., Oberwalder, C., Kabouw, P., Coulson, M., Ernst, G., Leicher, T., Miles, M., Weyman, G. & Klein, O. (2013): Occurrence and distribution of earthworms in agricultural landscapes across Europe with regard to testing for responses to plant protection products. J Soils Sediments 13: 278–293.

Dunger, W. & Fiedler, H-J. (1997): Methoden der Bodenbiologie. Fischer Verl. Jena. 539 pp.

Duquesne, S., Alalouni, U., Gräff, T., Frische, T., Pieper, S., Egerer, S., Gergs, R. & Wogram, J. (2020): Better define beta-optimizing MDD (minimum detectable difference) when interpreting treatment-related effects of pesticides in semi-field and field studies. Environ. Sci..Pollut. Res. 27: 8814 – 8821

EC (European Commission) (1991): Council Directive 91/414/EEC of 15 July 1991 concerning the placing of pesticides on the market, O.J. No L230, 19.8.1991 as amended by Directive 97/57/EC establishing Annex VI to Directive 91/414/EEC O.J. No. L265, 27.9.1997

EC (European Commission) (2002): Draft Working Document. Guidance Document on Ter-restrial Ecotoxicology Under Council Directive 91/414/EEC. SANCO/10329/2002 rev 2 final.

EC (European Commission) (2009): Regulation (EC) No 1107/2009 of the European parliament and the council of 21 October 2009 concerning the placing of pesticides on the market and repealing Council directives 79/117/EEC and 91/414/EEC. Off. J. Eur. Union L 309, 1–50.

EC (European Commission) (2013a): Commission Regulation No 283/2013 of 1 March 2013 setting out the data requirements for active substances, in accordance with Regulation (EC) No 1107/2009 of the European Parliament and of the Council concerning the placing of pesticides on the market. Brussels, Belgium

EC (European Commission) (2013b): Commission Regulation No 284/2013 of 1 March 2013 setting out the data requirements for pesticides, in accordance with Regulation (EC) No 1107/2009 of the European Parliament and of the Council concerning the placing of pesticides on the market. Brussels, Belgium.

EFSA (Panel on Pesticides and their Residues (PPR)). (2017): Scientific Opinion addressing the state of the science on risk assessment of pesticides for in-soil organisms. EFSA Journal 15:4690, 1-225 pp.

EFSA (European Food Safety Authority), 2019. Technical report on the outcome of the Pesticides Peer Review Meeting on general recurring issues in ecotoxicology. EFSA supporting publication 2019:EN-1673. 117 pp. doi:10.2903/sp.efsa.2019.EN-1673

Förster, B., Garcia, M.V.B., Oliveira, F. & Römbke, J. (2006): Effects of carbendazim and lambda-cyhalothrin on soil invertebrates and leaf litter decomposition in semi-field and field tests under tropical conditions (Amazonia, Brazil). Europ. J. Soil Biol. 42: S171-S179.

Fox D.R. (2009): Is the ECx a legitimate surrogate for a NOEC? Integrated Environmental Assessment and Management, 5: 351-353.

Horn, M.; Vollandt, R. (1995): Multiple Tests und Auswahlverfahren. Biometrie, Gustav Fischer, Stuttgart

ISO (International Organization for Standardization) (1999; 2014): Soil quality – Effects of pollutants on earthworms. Part 3: Guidance on the determination of effects in field situations. ISO 11268-3. Geneva, Switzerland.

Jager T. (2011): Some good reasons to ban ECx and related concepts in ecotoxicology. Environ. Sci. Technology, 45: 8180-8181.

Jager T. (2012): Bad habits die hard: The NOEC's persistence reflects poorly on ecotoxicology. Environmental Toxicology and Chemistry, 3182): 228-229.

James, S.W., Porco, D., Decaens, T., Richard, B., Rougerie, R. and Erseus, C. (2010). DNA barcoding reveals cryptic diversity in *Lumbricus terrestris* L., 1758 (Clitellata): Resurrection of *L. herculeus* (Savigny, 1826). PLoS ONE, Vol. 5 | Issue 12 | e15629.

Jänsch, S., Frampton, G.K., Römbke, J., Van den Brink, P.J. & Scott-Fordsmand, J.J. (2006): Effects of pesticides on soil invertebrates in model ecosystem and field studies: A review and comparison with laboratory toxicity data. Envir. Toxicol. Chem. 25: 2490-2501.

Jänsch, S., Steffens, L., Höfer, H., Horak, F., Roß-Nickoll, M, Russell, D., Toschki, A. and Römbke, J. (2013): State of knowledge of earthworm communities in German soils as a basis for biological soil quality assessment. Soil Organisms, 85: 215-232 + Electronic Supplement.

Jänsch, S., Alves, D., Cunha, L., Krogh, P.H., Natal-da-Luz, T., Rojo, V., Römbke, J., Sapkota, R., Scheffczyk, A., Schmelz, R.M., Scopel, L., Sousa, J.P., Vierna, J., Vizcaíno, A. (2025): Comparison of Morphological and DNA-Based Identification Methods to Assess Earthworm (Clitellata: Lumbricidae) Diversity at 25 Permanent Soil Monitoring Sites in Germany. https://doi.org/10.1002/ece3.71155

Knacker, T.H., Förster, B., Römbke, J. & Frampton, G. (2003): Methods for assessing the effects of pesticides on organic matter breakdown. Soil Biol. Biochem. 35: 1269-1287.

Knacker T, Van Gestel CAM, Jones SE, Soares AMVM, Schallnaß H-J, Förster B, Edwards CA. 2004. Ring-testing and field validation of a terrestrial model ecosystem (TME) – an instrument for testing potentially harmful substances: conceptual approach and study design. Ecotoxicology 13:5-23.

Kooijman S.A.L.M. (1996): An Alternative for NOEC exists, but the standard model has to be abandoned first. Oikos, 75(2): 310-316.

Krück, S. (2018): Bildatlas zur Regenwurmbestimmung. – Natur + Text, Rangsdorf. 196 pp.

Kula, C. & Römbke, J. (1998): Testing organic matter decomposition within risk assessment of pesticides. ESPR - Environmental Science & Pollution Research 5:55-60.

Kula, C., Heimbach, F. Riepert, F. & Römbke, J. (2006). Technical Recommendations for the update of the ISO Earthworm Field Test Guideline (ISO 11268-3). J Soils & Sed., 6: 182-186.

Landis W.G., Chapman P.M. (2011): Well past time to stop using NOELs and LOELs. Integrated Environmental Assessment and Management, 7(4): vi-viii.

Laskowski R. (1995): Some good reasons to ban the use of NOEC, LOEC and related concepts in ecotoxicology. Oikos, 73(1): 140-144.

Lavelle, P., Bignell, D. and Lepage, M. (1997): Soil function in a changing world: the role of invertebrate ecosystem engineers. Europ. J. Soil Biol., 33: 159-193.

Lavelle, P., Brussaard, L. & Hendrix, P. (1999): Earthworm management in tropical agroecosystems. CABI Publishing, Wallingford, UK, 300 pp.

Mulder, C., Boit, A., Bonkowski, M., De Ruiter, P.C., Mancelli, G., Van der Heijden, M.G.A., Van Wijnen, H.J., Vonk, J.A. & Rutgers, M. (2011): A belowground perspective on Dutch agroecosystems: how soil organisms interact to support ecosystem services. Adv. In Ecol. Res. 44:

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 (2005), *Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment*, OECD Series on Testing and Assessment, No. 34, OECD Publishing, Paris, https://doi.org/10.1787/e1f1244b-en.

OECD (2002), *Guidance Document for the Development of OECD Guidelines for Testing of Chemi*cals, OECD Series on Testing and Assessment, No. 1, OECD Publishing, Paris, https://doi.org/10.1787/9789264077928-en.

OECD (2006), *Guidance Document on the Breakdown of Organic Matter in Litter Bags*, OECD Series on Testing and Assessment, No. 56, OECD Publishing, Paris, https://doi.org/10.1787/a6204552-en.

OECD (2025), *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.

Pérez-Losada, M., Ricoy, M., Marshall, J.C. and Domínguez, J. (2009): Phylogenetic assessment of the earthworm *Aporrectodea caliginosa* species complex (Oligochaeta: Lumbricidae) based on mitochondrial and nuclear DNA sequences. Mol. Phylogen. Evol. 52: 293–302.

Petersen, H. & Luxton, M. (1982): A comparative analysis of soil fauna populations and their role in decomposition processes. Oikos: 39, 287-388.

Richard, B., Decaens, T., Rougerie, R., James, S.W., Porco, D. & Hebert, P.D.N. (2010): Re-integrating earthworm juveniles into soil biodiversity studies: species identification through DNA barcoding. Molec. Ecol. Res. 10: 606–614.

Römbke, J., Heimbach, F., Hoy, S., Kula, C., Scott-Fordsmand, J., Sousa, P., Stephenson, G. & Weeks, J. (eds.). (2003): Effects of Pesticides on Functional Endpoints in Soil (EPFES Lisboa 2002). SETAC Publ., Pensacola, USA. 92 pp.

Römbke, J., van Gestel, C.A.M., Jones, S.E., Koolhaas, J.E., Rodrigues, J.M.L. & Moser, T. (2004): Ring-Testing and Field-Validation of a Terrestrial Model Ecosystem (TME) – An Instrument for Testing Potentially Harmful Substances: Effects of Carbendazim on Earthworms. Ecotoxicology 13: 105-118.

Römbke, J., Jänsch, S., Murillo, L. & Sousa, J.P. (2010): Further development of a strategy for the assessment of the risk of pesticides to soil organisms – considering different life form types and exposure scenarios (FKZ 360 03 047). Report for the Federal Environmental Agency, Dessau-Roßlau. ECT Oekotoxikologie GmbH in cooperation with the University of Coimbra, Portugal. Unpubl. Report, 160 pp.

Römbke, J., Förster, B., Jänsch, S., Kaiser, F., Scheffczyk, A., Roß-Nickoll, M., Daniels, B., Ottermanns, R. & Scholz-Starke, B. (2020): Necessary adaptations for a harmonized field-testing procedure and risk assessment of earthworms (terrestrial). UBA-Texte. 193/2020, 118 pp.

Rutgers, M., Orgiazzi, A., Gardi, C., Römbke, J., Jänsch, S., Keith, A., Neilson, R., Boag, B. Schmidt, O., Murchie, A.K., Blackshaw, R.P., Pérès, G., Cluzeau, D., Guernion, M., Briones, M.J.I., Rodeiro, J., Piñeiro, R., Díaz-Cosin, D., Sousa, J.P., Suhadolc, M., Kos, I., Krogh, P.H., Faber, J., Mulder, C., Bogte, J.J., van Wijnen, H.J., Schouten, A.J. & de Zwart, D. (2016): Mapping earthworm communities in Europe. Applied Soil Ecology 97: 98-111.

Sherlock, E. (2018): Key to the earthworms of the UK and Ireland. FSC Publications Telford, UK. 92 pp.

Sims, R.W. & Gerard, B.M. (1999): Earthworms. In: Kermack, D.M. and Barnes, R.S.K. (eds): Synopses of the British Fauna (New Series), 31. 177 pp.

Stephan C.E., Rogers J.W. (1985): Advantages of using regression analysis to calculate results of chronic toxicity tests. In: Bahner R.C. and Hansen D.J (eds.): Aquatic toxicology and hazard assessment: Eighth symposium, ASTM STP 891. American Society for Testing and Materials, Philadelphia, pp. 328-338.

Stringer, A. & Wright M.A. (1973): The effect of Benomyl and some related compounds on *Lumbricus terrestris* and other earthworms. Pestic Sci 7:459–464.

Tanaka Y., Nakamura K., Yokomizo H. (2018): Relative robustness of NOEC and ECx against large uncertainties in data. PLOS ONE, 13(11): e0206901.

Toschki, A., Hammers-Wirtz, M., Poßberg, C., Roß-Nickoll, M., Schaeffer, A., Schmidt, B., Scholz-Starke, B., Römbke, J., Scheffzyk, A., Klein, M. & Hommen, U. (2018). Evaluation of the risk for soil organisms under real conditions - Development of a national position in the context of the new European Plant Protection Regulation (EU 1107/2009). UBA-Texte. (In press)

Van Groenigen, J.W., Lubbers, I.M., Vos, H.M.J., Brown, G.G., De Deyn, G.B. and Van Groenigen, K.J. (2014). Earthworms increase plant production: a meta-analysis. Scientific Reports 4: 6365 | DOI: 10.1038/srep06365. 7 pp.

Walter H., Consolaro F., Gramatica P., Scholze M., Altenburger R. (2002): Mixture toxicity of priority pollutants at no observed effect concentrations (NOECs). Ecotoxicology, 11: 299-310.

Warne M.S.J., van Dam R. (2008): NOEC and LOEC data should no longer be generated or used. Australasian Journal of Ecotoxicology, 14: 1-5.