SETAC PRESS

Environmental Toxicology and Chemistry, Vol. 35, No. 5, pp. 1297–1309, 2016
Published 2015 SETAC
Printed in the USA

Hazard/Risk Assessment

CRED: CRITERIA FOR REPORTING AND EVALUATING ECOTOXICITY DATA

CAROLINE T.A. MOERMOND,*† ROBERT KASE,‡ MURIS KORKARIC,§ and MARLENE ÅGERSTRAND‖
†Centre for Safety of Substances and Products, National Institute for Public Health and the Environment RIVM, Bilthoven, The Netherlands
‡Swiss Centre for Applied Ecotoxicology, Dübendorf, Switzerland
§Department of Environmental Toxicology, EAWAG, Dübendorf, Switzerland
‖Department of Environmental Science and Analytical Chemistry, Stockholm University, Stockholm, Sweden

(Submitted 5 June 2015; Returned for Revision 1 September 2015; Accepted 18 September 2015)

Abstract: Predicted-no-effect concentrations (PNECs) and environmental quality standards (EQSs) are derived in a large number of legal frameworks worldwide. When deriving these safe concentrations, it is necessary to evaluate the reliability and relevance of ecotoxicity studies. Such evaluation is often subject to expert judgment, which may introduce bias and decrease consistency when risk assessors evaluate the same study. The Criteria for Reporting and Evaluating Ecotoxicity Data (CRED) project attempts to address this problem. It aims to improve the reproducibility, transparency, and consistency of reliability and relevance evaluations of aquatic ecotoxicity studies among regulatory frameworks, countries, institutes, and individual assessors. In the present study, the CRED evaluation method is presented. It includes a set of 20 reliability and 13 relevance criteria, accompanied by extensive guidance. Risk assessors who participated in the CRED ring test evaluated the CRED evaluation method to be more accurate, applicable, consistent, and transparent than the often-used Klimisch method. The CRED evaluation method is accompanied by reporting recommendations for aquatic ecotoxicity studies, with 50 specific criteria divided into 6 categories: general information, test design, test substance, test organism, exposure conditions, and statistical design and biological response. An ecotoxicity study in which all important information is reported is more likely to be considered for regulatory use, and proper reporting may also help in the peer-review process. *Environ Toxicol Chem* 2016;35:1297–1309. © 2015 The Authors. *Environmental Toxicology and Chemistry* Published by Wiley Periodicals, Inc. on behalf of SETAC.

Keywords: Aquatic toxicology Reliability Relevance Study evaluation Reporting recommendation

INTRODUCTION

Ecotoxicity studies published in peer-reviewed literature contribute knowledge to the research community but also can be useful for regulatory purposes, such as for deriving predicted-no-effect concentrations (PNECs) for marketing authorizations or environmental quality standards (EQSs) to monitor environmental quality. However, evaluations of the reliability and relevance of (eco)toxicity studies often are subject to expert judgment. Even with an identical set of studies, risk assessors may not arrive at the same final list of reliable and relevant studies [1,2]. In hazard and risk assessment, a transparent evaluation process is needed, and clear documentation helps with understanding regulatory decisions. The use of ecotoxicity studies in regulatory risk assessments can be improved with the use of evaluation methods that guide risk assessors in performing unbiased, transparent, and detailed evaluations and with the use of reporting recommendations that guide researchers in performing and reporting studies that fulfill regulatory requirements. The present study addresses both issues.

Several chemical frameworks have recommended using the method of Klimisch et al. [3] for study evaluation, even though it has been found to be unspecific, to lack essential criteria and guidance for both reliability and relevance evaluations, and to leave considerable room for interpretation [4]. The Klimisch method has also been criticized for being biased toward interests of industry and for promoting use of guideline studies performed according to good laboratory practices (GLP) [5]. Altogether, this could result in a situation in which risk assessors arrive at different conclusions regarding the reliability and relevance of a study and whether it could be included in a specific regulatory process. Several improved evaluation methods exist [6,7], but none has replaced the Klimisch method in the European Union regulatory frameworks. The US Environmental Protection Agency (USEPA) has developed its own evaluation guidelines [8].

Regulatory assessments are often hampered by a lack of reliable (eco)toxicity studies, such as for nanoparticles [9,10], pharmaceuticals [11], and industrial chemicals [12–14]. Moreover, evaluations of recently published (eco)toxicity studies show incomplete and inadequate reporting, regarding both description of methodology and presentation of results [4,15]. Promotion of proper reporting by scientific journals has been suggested as a solution to this problem [16]. In several other research areas, systematic reporting recommendations have been developed to guide researchers, reviewers, and editors during the publication process, for instance, the STROBE statement in the field of epidemiology [17], the ARRIVE guideline for in vivo toxicity studies [18], *Nature*'s reporting checklist for life sciences articles [19], and the MIAME reporting standard for microarray experiments [20]. To our knowledge, no peer-reviewed journals currently use reporting recommendations for ecotoxicity studies. Implementation of reporting recommendations within the field of ecotoxicology may improve the reliability and reproducibility of studies and

This article includes online-only Supplemental Data.

* Address correspondence to caroline.moermond@rivm.nl

Published online 24 September 2015 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/etc.3259

streamline the publication process by ensuring that all essential information is provided.

CRED project: Criteria for reporting and evaluating ecotoxicity data

The CRED project aims to improve the reproducibility, consistency, and transparency of reliability and relevance evaluations of ecotoxicity studies, both within and between regulatory frameworks, countries, institutes, and individual assessors. Additional aims are to improve the usability of peer-reviewed literature for regulatory purposes and to facilitate the exchange of assessments between frameworks. Furthermore, the CRED project aims to improve the reporting of ecotoxicity studies by providing recommendations for reporting methodological details and results. The present study describes the CRED evaluation method, including extensive guidance on how to use the reliability and relevance criteria and it describes the CRED recommendations for reporting of ecotoxicity studies. The CRED project addresses aquatic ecotoxicity studies but can be adapted to other types of ecotoxicity studies.

METHOD

Within the CRED project, a ring test was performed in which the method of Klimisch et al. [3] was compared with the CRED evaluation method. Ring test participants were asked to evaluate 2 ecotoxicity studies each, selected from a total of 8 studies, using the Klimisch method (phase 1) and another 2 studies using the CRED evaluation method (phase 2). Ring test participants represented various institutions (industry, academia, consultancy, and governmental institutions) and geographical areas (Asia, Europe, and North America), and the majority had more than 5 yr of experience in study evaluation. More information about the ring test methodology and results can be found in the Supplemental Data, Appendix A, and in Kase et al. [21]. The CRED evaluation method was drafted based on existing evaluation and reporting methods [3,6,7,22–26] and on expert knowledge from the authors' respective institutions and ring test participants (after completing phase 1). The method was designed to fit to a broad range of aquatic ecotoxicity tests, from acute to chronic, with a variety of test organisms and test substances. After completing phase 2 of the ring test, the CRED evaluation method was fine-tuned using the participants' evaluation results and comments and input from discussions within 2 European risk assessment expert groups. For this fine-tuning process, focus was mainly on rewriting those evaluation criteria where the answers of the participants differed the most.

The CRED reporting recommendations were developed from the CRED evaluation method to facilitate reporting of studies with high reliability and reproducibility. The recommendations can serve as a template for supplemental data to peer-reviewed studies.

Excel files were created to facilitate the use of the CRED evaluation method and the CRED reporting recommendations by risk assessors and authors (Supplemental Data, Appendix B).

CRED EVALUATION METHOD

Reliability and relevance

Reliability and relevance are defined as follows [26]: "Reliability—evaluating the inherent quality of a test report or publication relating to preferably standardized methodology and the way the experimental procedure and results are described to give evidence of the clarity and plausibility of the findings"; "Relevance—covering the extent to which data and tests are appropriate for a particular hazard identification or risk characterization." From these definitions, it can be deduced that reliability concerns the intrinsic scientific quality of the study, regardless of the purpose for which it is assessed. On the other hand, relevance depends on the purpose of the assessment and concerns the way the study will be used for a specific purpose. Thus, a reliable study can be very relevant for one assessment but not relevant for another. Comments by ring test participants showed that relevance and reliability are used interchangeably and considered to be connected: "Can an unreliable study be relevant?" "A study must be reliable; otherwise, it cannot be relevant." "Is it necessary to assess relevance for an invalid study?" Care should be taken to avoid this mix-up, because reliability and relevance are 2 different aspects to be considered in the evaluation process.

Already when the literature is screened for ecotoxicity studies a rough assessment of the relevance of each study is performed based on title and abstract; for example, terrestrial ecotoxicity studies would be disregarded when an aquatic assessment is performed. The relevance evaluation described in the present study is primarily meant for a more in-depth evaluation, which takes place after the first selection of studies.

Reliability is determined by an assessment of the design, performance, and analysis of the experiment. A study may, for example, be considered less reliable because of an inadequate experimental design (e.g., too few replicates), poor performance (e.g., mortality is too high in the controls), or insufficient data analysis (e.g., inadequate statistics). An ecotoxicity study may contain several results (e.g., for mortality, reproduction, blood parameters) that may be obtained at different exposure scenarios. Within this single study, some results may be reliable and/or relevant, while others are not. In addition, a study with lower reliability and/or relevance may still be used in regulatory risk assessment as supporting evidence, depending on the reason for lowering the reliability/relevance. The following 3 examples illustrate the relation between reliability and relevance.

Example 1: A reliable toxicity study with a soil organism (e.g., Collembola) is not relevant for the derivation of water quality standards but may be relevant to derive soil quality standards or to assess the risk of a veterinary medicinal product for the terrestrial environment.

Example 2: An acute mortality study with fish may be reliable but not relevant for use in risk assessment of an endocrine disruptor because the mode of action of endocrine disruptors usually concerns chronic effects, which are not addressed in an acute toxicity study.

Example 3: A poster presenting study results on a sensitive test organism may not present enough information to determine its reliability. However, the results may be very relevant for a specific purpose and may serve as supporting evidence for regulatory decisions.

CRED reliability evaluation

The CRED evaluation method uses 4 reliability categories, similar to the Klimisch et al. scores [3]: reliable without restrictions (R1), reliable with restrictions (R2), not reliable (R3), and not assignable (R4). A more detailed description of these categories is provided in Table 1. The CRED reliability criteria are presented in Table 2.

In general, a study should only be assigned to be "reliable without restrictions" (R1) when all important information is provided and the study has no critical flaws in experimental

Table 1. Reliability categories[a]

Score	Description
R1	Reliable without restrictions: All critical reliability criteria for this study are fulfilled. The study is well designed and performed, and it does not contain flaws that affect the reliability of the study.
R2	Reliable with restrictions: The study is generally well designed and performed, but some minor flaws in the documentation or setup may be present.
R3	Not reliable: Not all critical reliability criteria for this study are fulfilled. The study has clear flaws in study design and/or how it was performed.
R4	Not assignable: Information needed to make an assessment of the study is missing. This concerns studies that do not give sufficient experimental details and that are only listed in abstracts or secondary literature (books, reviews, etc.) or studies of which the documentation is not sufficient for assessment of reliability for one or more vital parameters.

[a]Adapted from Klimisch et al. [3].

design and results. In contrast, "reliable with restrictions" (R2) can be assigned to studies in which not all details are given, raw data are not provided, or there are some minor flaws in experimental design, but for which it can still be assumed with reasonable certainty that the results are reliable.

It should be noted that assignment to "reliable without restrictions" (R1) and "reliable with restrictions" (R2) is not restricted to guideline and/or GLP studies. A properly performed and reported peer-reviewed study (whether GLP or not) may be evaluated as "reliable without restrictions" (R1), just as a poorly designed or performed guideline and/or GLP study should be assigned as "not reliable" (R3). "Not assignable" (R4) should be assigned if a study lacks the details necessary to evaluate reliability but is not per se "not reliable" (R3). In practice, however, guideline and/or GLP studies often contain all details necessary to evaluate reliability and will not be assigned to be "not assignable" as often as studies reported in peer-reviewed literature. A prerequisite to enable a thorough evaluation is proper reporting of the methods used and the results obtained. However, transparent reporting is not in itself critical for evaluating reliability but does allow for fast and easy review of the data. Thus, a study should be evaluated based on the details provided rather than how well the report is written, unless the writing is so poor that the methods are not clear. Providing only a statement that a study was performed according to a certain guideline is not detailed enough, because

Table 2. CRED reliability criteria[a]

Number	Criterion
General information	
	Before evaluating the test, check the physicochemical characteristics of your compound (handbooks/general sources). What is the solubility, log K_{OW}, or pK_a? Is the compound volatile? Does it hydrolyze, photolyze, etc.?
Test setup	
1	Is a guideline method (e.g., OECD/ISO) or modified guideline used?[b]
2	Is the test performed under GLP conditions?[b]
3	If applicable, are validity criteria fulfilled (e.g., control survival, growth)?
4	Are appropriate controls performed (e.g., solvent control, negative and positive control)?
Test compound	
5	Is the test substance identified with name or CAS number? Are test results reported for the appropriate compound?
6	Is the purity of the test substance reported? Or, is the source of the test substance trustworthy?
7	If a formulation is used or if impurities are present: Do, other ingredients in the formulation exert an effect? Is the amount of test substance in the formulation known?
Test organism	
8	Are the organisms well described (e.g., scientific name, weight, length, growth, age/life stage, strain/clone, gender if appropriate)?
9	Are the test organisms from a trustworthy source and acclimatized to test conditions? Have the organisms not been pre-exposed to test compound or other unintended stressors?
Exposure conditions	
10	Is the experimental system appropriate for the test substance, taking into account its physicochemical characteristics?
11	Is the experimental system appropriate for the test organism (e.g., choice of medium or test water, feeding, water characteristics, temperature, light/dark conditions, pH, oxygen content)? Have conditions been stable during the test?
12	Were exposure concentrations below the limit of water solubility (taking the use of a solvent into account)? If a solvent is used, is the solvent within the appropriate range and is a solvent control included?
13	Is correct spacing between exposure concentrations applied?
14	Is the exposure duration defined?
15	Are chemical analyses adequate to verify concentrations of the test substance over the duration of the study?
16	Is the biomass loading of the organisms in the test system within the appropriate range (e.g., <1 g/L)?
Statistical design and biological response	
17	Is a sufficient number of replicates used? Is a sufficient number of organisms per replicate used for all controls and test concentrations?
18	Are appropriate statistical methods used?
19	Is a concentration–response curve observed? Is the response statistically significant?
20	Are sufficient data available to check the calculation of endpoints and (if applicable) validity criteria (e.g., control data, concentration–response curves)?

[a]See main text for further explanation of the criteria and explanatory guidance text on how to interpret the criteria. Please note that most criteria are not per se critical for the reliability of a study and that this depends strongly on the compound and/or species tested.
[b]These criteria are of minor importance for study reliability but may support study evaluation.
CRED = criteria for reporting and evaluating ecotoxicity data; ISO = International Organization for Standardization; GLP = good laboratory practice; CAS = Chemical Abstracts Service; K_{OW} = octanol–water partition coefficient; OECD = Organisation for Economic Co-operation and Development; pK_a = dissociation constant.

specific information on the experimental design is necessary to decide if the performed test is suitable for the substance and organisms tested. Thus, a study that was carried out in a scientifically sound way may be assigned to be "unassignable" if the description of the method lacks details (e.g., experimental setup is given as a reference to another report that cannot be retrieved) or if criteria that are considered important for interpretation of the test results cannot be evaluated (e.g., temperature data are not given) and these data have not been retrieved by the assessor [15,24]. If necessary and possible, the authors of a specific study can be asked to provide the details needed. However, when there are clear flaws in the study setup or the results, additional information about the study will not remove these flaws and the study should be assigned to "not reliable" (R3). The CRED ring test showed that the categories "not reliable" (R3) and "not assignable" (R4) were often mixed up; ring test participants misunderstood R4 and wrongly assigned the R3 category to studies when information was missing but no flaws in the design and performance were apparent. Care should be taken to avoid this misinterpretation.

Explanation of the reliability criteria (criteria numbers from Table 2)

Physicochemical parameters. Knowledge on physicochemical parameters of the substance under investigation and its behavior in relevant environmental compartments is essential when deciding on the reliability of a study setup. For instance, solubility in water should not be exceeded, different exposure conditions are needed for volatile substances than for hydrophobic substances, and the possible occurrence of hydrolysis should be taken into account. These parameters do not need to be given in the study, since often they can be found in chemical handbooks, pesticide manuals, and so forth.

1. Is a guideline method (e.g., Organisation for Economic Co-operation and Development [OECD], International Organization for Standardization [ISO]) or modified guideline used? Use of a guideline method (OECD, ISO, USEPA, or comparable) does not necessary reflect the reliability of a study, and it should therefore never be a critical criterion. A guideline study may be unreliable if there are flaws in the design, conduct, and/or (statistical) interpretation or if results give rise to doubt. This may occur when, for instance, exposure conditions are not suitable for the substance under investigation, control mortality is too high or other validity criteria are not met, or when presumed outliers are left out of consideration without proper justification. The criterion is included in the CRED evaluation method to aid with transparent evaluation. Studies using nonguideline methods may be equally reliable as guideline studies, provided enough details on the experimental design and results are presented to assess its reliability.

2. Is the test performed under GLP conditions? Good laboratory practice is a data quality system that requires adequate documentation of the experimental process. Laboratories working under GLP often use standardized methods, both for performing the test and for documenting the results. Good laboratory practice does not, however, reflect the actual reliability of a study and should therefore never be a critical criterion. It is included in the CRED evaluation method to aid with transparent evaluation.

3. If applicable, are validity criteria fulfilled (e.g., control survival, growth)? In most test guidelines, validity criteria are provided to determine the validity of the test results. For instance, OECD guideline 201 on algal toxicity requires exponential growth in the controls and specifies criteria for the variation in growth rate within and between control replicates. For the *Daphnia* acute toxicity study, the validity criteria in the OECD 202 guideline include control mortality and oxygen concentrations. Besides this, control organisms should be from the same population as the treatment group(s), variability in the controls should fall within the same range as historical data, and attention should be given to natural fluctuations in results, such as fluctuations attributable to age of the animals or seasonal influences. If a nonguideline test is performed with a guideline species, validity criteria as described in the relevant guideline should be met. If nonguideline species are used, expert judgment is needed to assess whether the test organism resembles the guideline test species enough to apply guideline validity criteria. Otherwise, expert judgment is needed to decide if control survival and/or other parameters are within the range of what is normal for the species and that other confounding (stress) factors can be ruled out. For guideline test species, however, complying with guideline criteria for validity (e.g., control survival, growth) is critical for a study to be reliable.

4. Are appropriate controls performed (e.g., solvent control, negative and positive controls)? The decision of which controls to use depends on the test substance and/or the guideline applied. Next to the "normal" negative controls (no solvent, no test substance), solvent controls need to be tested in all cases where a solvent is used. The concentration of solvent in the solvent control should be the same as the highest concentration used in the test treatments, and mortality in the solvent controls should preferably not differ significantly from that in the nonsolvent controls. The study is "not reliable" (R3) if, for instance, a solvent is used but no solvent control is tested or the solvent concentration in the control treatment is too low. If the mortality in the solvent control is higher than that in the nonsolvent control, statistics should be based on the solvent control. Expert judgment is needed to decide when the mortality in the solvent control is too high, especially when it is still within the validity criteria of the test.

In some cases, a positive control (with a reference substance) is tested. The lack of a positive control decreases the reliability of a study only if a positive control is requested in the test guideline. On the other side, use of a positive control might increase confidence regarding the reliability of study results.

5. Is the test substance identified clearly with name or Chemical Abstracts Service (CAS) number? Are test results reported for the appropriate substance? It is essential to know which substance was tested. If a salt was tested, for example, information on the type of salt and how results are reported (e.g., as salt or as its positively or negatively charged ions, including or excluding waters of hydration as $\cdot H_2O$) is needed. The only exception for this is when results are expressed in molarities instead of grams per liter. In that case, it does not matter if results are expressed as salt or base; although it remains necessary to know which salt is tested, because different counter-ions may give different results, or if a salt contains more than one of the same ion. The lack of a CAS number does not decrease the reliability of the study since it can often be retrieved easily from the Internet or other sources if the tested substance is clearly identified. If a formulation is tested, it is necessary to know all the relevant components of the formulation (see also criterion no. 7).

Example 4: A test with metformin hydrochloride is performed. Values for the no-observed-effect concentration (NOEC) and 50% effect concentration (EC50) are tabulated for "metformin," but it is not specifically mentioned if results are reported for the metformin base or the salt. If metformin

hydrochloride was the tested substance, the toxicity values for metformin base would be lower. This decreases the reliability of the study, although results may still be used as supporting evidence in risk assessment.

6. Is the purity of the test substance reported? Or, is the source of the test substance trustworthy? The purity of the substance and/or the source of the substance should be reported and reliable (e.g., a known supplier). Generally, a substance should have a purity of 80% or higher, unless it is known that the impurities do not cause toxic effects and do not influence the toxicity of the substance of interest. When the purity of the substance is <90%, the nominal test results should be corrected for purity [27]. Although the OECD guidelines only refer to a "suitable purity," the overall reliability of a study that uses a low-purity substance should be lowered. However, when the purity is not known but actual test concentrations are measured, this criterion becomes less important.

7. If a formulation is used or if impurities are present: Do, other ingredients in the formulation exert an effect? Is the amount of test substance in the formulation known? If a formulation is used for testing a specific compound (e.g., when testing plant protection products), the other constituents of the formulation should be known and/or it should be clear that these other constituents have no ecotoxicological effects. For a study to be reliable, the amount of active substance in the formulation should be known, and it should be clear that the results are (or can be) expressed in terms of active substance. It should be taken into account that certain "inert" ingredients of formulations can exert biological effects, because coformulants and solvents in formulations may significantly increase or decrease the toxicity of the active substance and there is some difficulty in predicting which type of formulations are critical in terms of such interactions [28]. However, this information might be confidential and, for that reason, not known. If this is the case, the reliability should be lowered. In contrast to this, when the formulation itself (and not the active ingredient only) is evaluated for hazard or risk assessment, it may be enough information to have the exact name of the formulation, including formulation strength, and results may be expressed in amounts of formulation.

8. Are the organisms well described (e.g., scientific name, sex, weight, length, growth, age/life stage, strain/clone, gender, if appropriate)? When assessing reliability, it is essential to know which organisms were used in the test. At a minimum, the name and information on age or life stage should be known for a study to be reliable. Other information such as weight, length, or strain/clone is in most cases not essential for the reliability assessment, but it may increase the confidence in the study. When examining hormonal substances, the sex of the organisms (e.g., when testing fish) may influence the results, and this information should thus be known.

Example 5: When performing a test with fish, larvae and juveniles are often more sensitive than adults. Thus, when evaluating the test or when comparing results from different tests, knowledge on the life stages tested may be important.

9. Are the test organisms from a trustworthy source and acclimatized to test conditions? Have the organisms not been pre-exposed to the test substance or other unintended stressors? The source of the test organisms should be known and trusted, and the place of origin should be described for field-collected organisms. Test organisms should be healthy and acclimatized to test conditions (e.g., water type, temperature) to avoid any unintentional stress caused by a change in conditions. In addition, organisms should not be stressed by test conditions (except when this is part of the research question) or by other (unintended) stressors. When such stress is reflected in high control mortality, the study becomes unreliable. Results of toxicity tests with field-collected test organisms may be biased because of community adaptation to toxic stress if pre-exposure to the test substance has occurred [29]. If this is the case, the study becomes unreliable.

10. Is the experimental system appropriate for the test substance, taking into account its physicochemical characteristics? Most test guidelines prescribe or give recommendations for the experimental system but allow for flexibility with respect to the actual design. Some requirements depend on the substance and/or organism used. The demands of the appropriate test guidelines should be followed as closely as possible, also for nonguideline test organisms. For instance, the test vessel should preferably be made of glass, but this demand is more important for some substances than for others. When hydrophobic substances are tested in paper cups or plastic containers, for example, the study becomes unreliable because of sorption of the substance to the test vessel.

Static systems may be appropriate for short-term tests with stable substances. However, static systems are usually not appropriate for long-term exposure. Open systems may be used for most substances, but volatile substances need to be tested in a closed test system for the study to be reliable, unless chemical analyses show that volatilization has not occurred during the experiment. Regular analysis of test concentrations can confirm maintenance of exposure concentrations during the test or be used to calculate actual exposure if concentrations declined. If test concentrations have not been stable during the test or no measurements were performed, it should be clear from the study report that all possible measures have been taken to avoid loss. If this is the case, the study can still be reliable. If not, the study is "not reliable" (R3); or, if not enough details are provided, the study is "not assignable" (R4). As indicated before, this criterion may not be applicable to substances that are known to be stable in solution.

Example 6: Unfiltered natural water is used for an acute Daphnia toxicity study to test a substance with log octanol–water partition coefficient (K_{OW}) of 4.2. Exposure concentrations are measured in unfiltered water. No information is provided on the amount of particulate matter in the water. Because of the high sorption of the substance, it can be assumed that a significant amount of the substance sorbs to the particulate matter. For many substances, it is assumed that the toxicological effect is caused by the dissolved fraction. Thus, the actual exposure concentration is not known, and the study is less reliable. In contrast, if dissolved concentrations are measured and the total organic carbon is below 2 mg/L (OECD 202 requirement), the study is reliable.

11. Is the experimental system appropriate for the test organism (e.g., choice of medium or test water, feeding, water characteristics, temperature, light/dark conditions, pH, oxygen content)? Have conditions been stable during the test? The experimental system should be appropriate for the test organisms; for instance, freshwater species should not be tested in salt water. However, which test conditions are considered to be appropriate depends on the organism tested, and no specific guidance can be given. When testing a photodegradable substance, for example, the experiment may be performed under dark conditions for fish and daphnids, but algae will need light to grow. Temperature, pH, and oxygen should preferably be stable and within the appropriate range for the organism and the substance. If there is a large variability among the controls or

the control performance is not good (e.g., high mortality), this may indicate that the test conditions were not appropriate and the study is not reliable.

Feeding is not allowed in acute toxicity studies because of interference with the test substance. For chronic studies, however, feeding is often necessary to keep the animals alive. Feeding should then follow the requirements of the guideline (if applicable), and all excess food should be carefully removed shortly after feeding to avoid decreased bioavailability of the test substance.

Sometimes modified exposure studies with sediment are performed with species that are normally tested in water-only systems. Endpoints from these studies will mostly be considered unreliable, especially when hydrophobic substances are tested, unless concentrations in the water phase are adequately monitored and reported. For aquatic insects that need some kind of substrate when tested chronically, inert substrates such as quartz sand or glass beads may be an appropriate alternative according to OECD guideline 233.

Example 7: An ecotoxicity study with a freshwater fish is performed in 3 different media: distilled water, reconstituted water, and water with different salinities. The controls show high mortality in the tests performed in distilled water and high salinity but no mortality in the tests performed in reconstituted water and low salinity. This implies that the endpoints from the tests in distilled water and high salinity are not reliable because of the stress these test media caused. However, the endpoints from the tests in reconstituted water and low salinity, where the fish only experienced the stress from the substance under investigation, are reliable.

12. Were exposure conditions below the limit of water solubility (taking the use of a solvent into account)? If a solvent is used, is the solvent within the appropriate range and is a solvent control included? If substances are tested at concentrations below the water solubility, the test can be assumed to be reliable. Depending on the uncertainty in the estimate of the water solubility, results of tests performed above the estimated water solubility may be reliable as well. Expert judgment should be used here. Reports of precipitates may indicate that the solubility was exceeded. In this case, test results are less reliable since actual concentrations do not equal nominal concentrations.

Results from a test in which a substance was tested at nominal concentrations 10 times higher than the solubility should normally be regarded as "not reliable" (R3) [24]. For tests where this is not the case, the reliability should be subject to expert judgment.

Solvents may be used to prepare stock solutions. It is, however, usually not advised that solvents be added directly to the test vessels to enhance solubility. Solvents should not be toxic to the tested species at the tested concentrations [26]. According to several OECD guidelines, the concentration of solvents should not exceed 0.01%. The highest solvent concentration used should be reflected in a solvent control (see also criterion number 5).

13. Is correct spacing between exposure concentrations applied? When spacing between test concentrations is too large, the results are not reliable, especially when deriving a NOEC value. A scaling factor of 3.2 ($=\sqrt{10}$) is often recommended. As a rule of thumb, a maximum scaling factor of 10 should be applied. Performing a range-finding test may help in determining the right exposure concentrations, and as a result the necessary spacing between exposure concentrations may be reduced.

Example 8: A limit study is performed at 2 concentrations: 0.01 mg/L and 1 mg/L. No effect was observed at 0.01 mg/L, and all organisms were dead at 1 mg/L. The NOEC value could thus have been 0.01 mg/L but also 0.1 mg/L or 0.25 mg/L. Because of this large spacing between the 2 concentrations and the low number of concentrations tested, it is not known what the actual effect concentration is. Thus, no reliable NOEC, EC50, or lowest-observed-effect concentration values can be derived from this study. However, the study may be used as supportive evidence, for example, to demonstrate the (in) sensitivity of the particular species in relation to other data.

14. Is the exposure duration defined? The exposure duration should be defined for a study to be reliable. Especially when results from different studies with the same test species are compared, it is necessary to know if the exposure durations were similar. The ideal test duration depends on the test organism used and the endpoint under investigation. Sometimes the exposure is very short (e.g., 1 d), whereas effects are not observed until days or weeks after exposure (delayed effects). Results should then be expressed in terms of the actual exposure duration and not in terms of the duration of the entire experiment, although the delay of the effects should be clearly mentioned when summarizing the results to enable comparison with other studies in which the observation of effects was stopped immediately after exposure.

15. Are chemical analyses adequate to verify concentrations of the test substance over the duration of the study? It is important to know the actual exposure concentrations, and it should be clear if the reported concentrations are initial or final concentrations, whether they are mean or geometric mean concentrations, and which of these concentrations are used to calculate the effect concentrations. In some cases, such as acute toxicity tests or semistatic (renewal) chronic tests with a stable substance, nominal concentrations without measurements can be acceptable. A static or semistatic acute study with a stable substance (information on stability should then be available from other experimental work or from physicochemical characteristics) may be reliable if no measurements are performed; but in all other cases, the exposure concentrations should be verified by analytical measurements. Analysis of the concentration at the beginning of the test may be enough, depending on the substance and the test system, but measurements usually should be performed at least at the beginning and the end of the experiment. There should be no major loss as a result of degradation, photolysis, volatilization, hydrolysis, or adsorption to glass or other equipment. During the experiment, the test concentration should be close to nominal (80–120%), especially when possible loss mechanisms are unknown, and the test design should be adequate to maintain this concentration. If the loss of test substance is higher than 20%, it should be investigated whether this is caused by insufficient performance of the test (in which case the reliability is reduced) or by other loss processes. Since these loss processes may be caused by the intrinsic properties of a substance, it may be impossible to avoid them. In case hydrolysis may have occurred, calculations could be performed by a quantitative structure–activity relationship specialist if experimental data on hydrolysis are missing. If test concentrations deviate more than 20% from nominal, it should be clear that all possible measures have been taken to maintain test concentrations (e.g., renewal, flow-through system). In this case, when the experiment is assessed to have been performed in a technically adequate way, the test could be considered "reliable with restrictions" (R2) [24]. Reported test results should then be based on measured concentrations, preferably as

time-weighted averages. However, it should be noted if metabolites were present, and expert judgment is needed to decide whether or not it is most suitable to express effects using the concentrations of the parent substance.

The method used to perform chemical analyses should preferably be reported. If the limit of detection and recovery efficiency of the method used are reported, the reliability of the study increases. However, lack of information on recovery and limit of detection does not make the study unreliable.

16. Is the biomass loading of the organisms in the test system within the appropriate range (e.g., <1 g/L)? Especially for hydrophobic substances, organism loading should be taken into account to avoid loss of the test substance by sorption to biota. This is mainly relevant for studies with larger organisms, like fish and macrophytes. Sorption to biota may become relevant when testing substances with log K_{OW} values >3. In addition, density stress may interfere with the effects of the chemical substance. The OECD guideline for acute fish toxicity tests recommends a maximum loading of 1.0 g fish/L for static and semistatic tests. For flow-through systems, higher biomass loadings may be acceptable if this does not cause a decrease in concentration of the test substance due to sorption to biomass.

Example 9. An experiment is performed using 30-L aquaria containing 40 fish of 1.5 g each (biomass loading of 2 g/L). If the experiment is static, the biomass loading would be too high. However, when a flow-through study is performed with a renewal rate of, for example, 90 L/d, the biomass loading would be acceptable.

17. Is a sufficient number of replicates used? Is a sufficient number of organisms per replicate used for all controls and test concentrations? In general, the guideline requirements for number of replicates should be used. When a nonguideline study is evaluated, expert judgment is needed to assess whether the study design is appropriate to obtain statistically reliable results. Statistically significant results do not automatically mean that the study is reliable, especially when there have been flaws in the study setup or in the performance of the study. For example, the use of pseudoreplicates (see example 10) lowers the reliability.

Example 10: An experiment is performed with 4 different concentrations and a control. For this experiment, 5 aquaria are used (1 per concentration). Within the aquaria, 3 compartments are made to separate groups of animals. These 3 groups of animals are considered to be pseudoreplicates (they have the same container, water, light, and temperature conditions and could experience the same stress). Thus, because the test is performed with pseudoreplicates only, the reliability is lowered.

18. Are appropriate statistical methods used? In general, the guideline requirements for statistics should be followed and a description of the statistics is needed to assess the reliability of an endpoint. When a nonguideline study is evaluated, expert judgment may be needed.

If effect values are missing, concentration–response data reported in tables or graphs can be used to calculate them. For example, the concentration at which 10% effect is observed (EC10) can usually be calculated when a concentration–response curve is available, and computer programs are available to translate graphs into individual data points (e.g., Techdig). However, effect values should be determined by interpolation and not by extrapolation, and they should preferably not exceed or be lower than the tested concentrations. A calculated EC10 value that is more than 3 times lower than the lowest tested concentration is less reliable. An NOEC value should be determined using an appropriate statistical method and should not be determined by visual inspection of the graphs or other estimation without statistical significance being determined.

19. Is a concentration–response curve observed? Is the response statistically significant? The requirement for a concentration–response relationship depends on the objective of the study. If an effect needs to be demonstrated, a concentration–response relationship is needed. However, when the study has been performed to verify that there is no effect at a certain dose, a concentration–response relationship is not necessarily needed to derive an NOEC value. Generally, if no monotonic concentration–response curve is observed, it is difficult to obtain reliable endpoints. Exceptions occur, for instance when increased growth is observed at low concentrations and a toxic effect is observed at high concentrations. The concentration–response curve can then be difficult to calculate, and a calculation model that allows for these effects is needed. When limit tests (with just 1 or 2 concentrations) are performed, no concentration–response curve can be observed. However, a properly designed limit test based on range-finding data and conducted at the limit of solubility is reliable, as long as no adverse effects are observed. If adverse effects do occur, then the study alone cannot be used to calculate a safe concentration.

If tables or graphs (which can be transformed back into numbers) with concentration–response data are available, effect values may be calculated. However, calculating a statistically significant NOEC value will not be possible if raw data or standard errors are missing. Nevertheless, it may then be possible to calculate an EC10 value, with statistical significance values, using programs such as GraphPad Prism. If endpoints with their statistical method are provided but no concentration-response graph or table is reported, the study can still be assigned to be "reliable with restrictions" (R2).

20. Are sufficient data available to check the calculation of endpoints and (if applicable) validity criteria (e.g., control data, concentration–response curves)? The availability of raw data is not a prerequisite for a study to be reliable. However, where "reliable with restrictions" (R2) can be assigned if raw data are not reported, "reliable without restrictions" (R1) can only be assigned when raw data are provided. By "raw data" we mean the data needed to assess the statistics and variability in the controls, recalculate the reported endpoints, and calculate alternative endpoints. These data may be presented in the form of tables or graphs, in the publication itself, or, in case of peer-reviewed papers, in the supplemental data.

Relevance evaluation

In contrast to reliability, relevance does not concern the inherent quality of the study but mainly depends on the purpose of the assessment or regulatory framework for which it is evaluated. Thus, relevance may change depending on the use of the study. For instance, a sediment toxicity study can be irrelevant for aquatic EQS or PNEC derivation but very relevant for risk assessment for sediment. This implies that most relevance aspects can be evaluated only if the framework and the purpose for the risk assessment are known.

Similar to the criteria for reliability, a distinction can be made between relevant and nonrelevant studies. The CRED evaluation method uses 4 relevance categories: relevant without restrictions (C1), relevant with restrictions (C2), not relevant (C3), and not assignable (C4). A more detailed description of these categories is provided in Table 3. A similar assignment of

Table 3. Relevance categories

Score	Description
C1	Relevant without restrictions: The study is relevant for the purpose for which it is evaluated.
C2	Relevant with restrictions: The study has limited relevance for the purpose for which it is evaluated.
C3	Not relevant: The study is not relevant for the purpose for which it is evaluated.
C4	Not assignable: Studies that do not give sufficient details since the result is presented in abstracts or secondary literature (books, reviews, etc.) or studies of which the documentation is not sufficient for assessment of relevance for one or more vital parameters.

these categories can be applied as explained for reliability. The relevance criteria are summarized in Table 4. Each criterion is further explained below.

1. Is the species tested relevant for the compartment under evaluation? The species tested should be relevant for the compartment under evaluation. For aquatic ecotoxicity studies, the test species should be relevant for the aquatic compartment. For instance, soil organisms such as nematodes, even when tested in aqueous medium, have lower relevance for aquatic risk assessments. Likewise, terrestrial plants could give information on the sensitivity of plants to a substance but would be of lower relevance in aquatic risk assessments. Depending on the substance and the framework, saltwater species may or may not be relevant for a freshwater assessment and vice versa.

2. Are the organisms tested relevant for the tested substance? Because the purpose of most assessments is to evaluate the potential risks of a substance to sensitive nontarget organisms, care should be given to the representativeness of test species; for example, an insecticide should preferably also be tested on insects, and an antimicrobial substance on cyanobacteria. For a study to be relevant, the test organisms do not necessarily have to be a test species for which an accepted test guideline is available. Information from nonsensitive species can also be relevant, especially when enough data are available to perform a species sensitivity distribution or for hazard assessments. When endocrine-disrupting substances are tested, effects might differ between males and females; for example, one substance mainly affects egg production, and another substance only affects sperm viability and fertility. Thus, a distinction between data on male and female organisms should be made for these kinds of compounds since the relevance of study results could differ between sexes.

Example 11: An insecticide is tested on algae. Although these are not a potentially sensitive species group for this substance, algal growth inhibition is still relevant for PNEC or EQS derivation and is needed for the "base set" of algae, invertebrates, and fish that is required in most regulatory frameworks.

3. Are the reported endpoints appropriate for the regulatory purpose? For PNEC and EQS derivation, studies on bioaccumulation may not be relevant. For the determination of an acute EQS, chronic data may be less relevant and vice versa.

4. Are the reported endpoints appropriate for the investigated effects or the mode of action of the test substance? When a risk assessment is performed for substances with a specific mode of action or a known adverse outcome pathway, studies that assess this particular mode of action or adverse outcome pathway are most relevant. For example, fish biomarkers, vitellogenin concentrations, secondary sex characteristics, and sex ratio are considered to indicate endocrine-disrupting chemicals interfering with estrogens, androgens, and steroidogenesis pathways [30]. These biomarkers, however, are not useful for indicating other modes of action such as the glucocorticoid receptor pathway. However, even if the use of a biomarker is not (yet) accepted for use in EQS derivation [31], studies on this biomarker can still be listed as supporting information in dossiers, to show the concentration range in which effects may occur.

Table 4. CRED relevance criteria[a]

Number	Criterion
General	
	Before evaluating the test for relevance, indicate why you are evaluating this study. The relevance of the study might be different for different purposes (e.g., environmental quality criteria derivation, PBT assessment, dossier evaluation for marketing authorization), also depending on the framework for which the evaluation is requested.
Biological relevance	
1	Is the species tested relevant for the compartment under evaluation?
2	Are the organisms tested relevant for the tested compound?
3	Are the reported endpoints appropriate for the regulatory purpose?
4	Are the reported endpoints appropriate for the investigated effects or the mode of action of the test substance?
5	Is the effect relevant on a population level?
6	Is the magnitude of effect statistically significant and biologically relevant for the regulatory purpose (e.g., EC10, EC50)?
7	Are appropriate life stages studied?
8	Are the experimental conditions relevant for the tested species?
9	Is the exposure duration relevant and appropriate for the studied endpoints and species?
10	If recovery is studied, is this relevant for the framework for which the study is evaluated?
Exposure relevance	
11	In case of a formulation, other mixture, salts, or transformation products, is the substance tested representative and relevant for the substance being assessed?
12	Is the tested exposure scenario relevant for the substance?
13	Is the tested exposure scenario relevant for the species?

[a]See main text for further explanation of the criteria and explanatory guidance text on how to interpret the criteria.
CRED = criteria for reporting and evaluating ecotoxicity data; PBT = persistent, bioaccumulative, and toxic; EC10/EC50 = 10% and 50% effect concentrations.

5. Is the effect relevant on a population level? Most frameworks consider only traditional test endpoints, such as mortality, growth, and reproduction, which are assumed to be linked to population sustainability. However, nonguideline tests may also report nonguideline or nonstandard endpoints that could be relevant, such as filtration rate and behavioral endpoints. The discussion on which endpoints are population-relevant is ongoing and differs between frameworks. Examples of debated endpoints include blood parameters, general behavior, swimming speed, gene expression, vitellogenin concentrations, in vitro tests, and coloration.

6. Is the magnitude of effect statistically significant and biologically relevant for the regulatory purpose (e.g., EC10, EC50)? In a standardized test system with relatively little control variation, minor changes may be statistically significant without necessarily being considered ecologically relevant. Expert judgment is needed to decide if the observed effect is caused by the chemical under investigation, especially when no concentration–response relationship is observed. Please note that if enough data are presented in tables or graphs, additional endpoints may be calculated by the assessor if not reported in the study.

For the derivation of chronic risk limits, EC10 and NOEC values are the preferred type of effect values. However, EC50 values can be used if EC10 or NOEC values are missing. If in a certain data set an EC50 value from an acute study is lower than the lowest NOEC value from chronic studies, this information is relevant for the risk assessment. For the derivation of acute risk limits in the European Union, EC50 values are preferred and NOEC/EC10 values derived from acute studies are less relevant.

Example 12: An NOEC value is available from an acute toxicity study with *Gammarus* sp. This NOEC value is well below the lowest available EC50 values for other species and, thus, indicates that *Gammarus* sp. is a very sensitive species. Although in the European Union, EQSs or PNECs for acute toxicity are not based on NOECs, depending on the regulatory framework this information might be used to adjust the assessment factor. In contrast, if the acute NOEC value were higher than EC50 values for other species, this would indicate that *Gammarus* sp. is not sensitive to this substance.

7. Are appropriate life stages studied? The studied life stage should be appropriate for the experimental design and the purpose of the study. For instance, an early life stage test with fish embryos or larvae is relevant for investigations of developmental effects but not relevant for investigating effects on reproduction.

8. Are the experimental conditions relevant for the tested species? Not only the species (criterion 1) but also the exposure route and conditions should reflect the compartment under investigation. For instance, freshwater species should be tested in freshwater, and saltwater species should be tested in salt water. If this is not the case, the result may not be relevant. If organisms are exposed through water (e.g., *Chironomus* sp.) and sediment is only needed to provide hiding space or as substrate for eggs, inert alternatives such as glass beads, silica sand, and cotton sheets may be used to prevent interference with the substance in the water phase.

9. Is the exposure duration relevant and appropriate for the studied endpoints and species? The exposure time should be in line with the endpoints and the test organism under investigation. For algae, the maximum exposure time is usually 3 d to 4 d; but depending on the test species, 7-d exposure may also be used. Although most guidelines recommend exposure for 96 h for an acute toxicity tests on fish, this does not mean that a 5-d or 10-d test is not relevant. Expert judgment is needed to decide whether a test should be considered acute or chronic. When studying chronic effects, sensitive life stages should be included or a whole life cycle should be studied.

10. If recovery is studied, is this relevant for the framework for which the study is evaluated? Recovery is not taken into account in most frameworks, the exception being the European authorization of plant protection products, where results based on recovery are relevant for risk assessments.

11. In case of a formulation, other mixture, salts, or transformation products, is the substance tested representative and relevant for the substance being assessed? A substance may be tested as a pure active substance or in a formulation. Tests performed with formulations may be of lower relevance for EQS derivation within the Water Framework Directive and of higher relevance for assessments within the Plant Protection Product framework. For pharmaceuticals, the metabolite excreted by humans or livestock may be more relevant for risk assessment than the parent substance. For unstable substances, it should be known if transformation products are formed and if these transformation products are toxic. If the substance causing the effect is not the substance under investigation, expert judgment is needed to decide on how to deal with the results of the study and the resulting risk assessment.

12. Is the tested exposure scenario relevant for the substance? The exposure scenario includes duration of exposure, exposure concentrations, application of the substance, route of administration, and the exposure schedule (static, semistatic, renewal, flow-through, etc.). Some exposure scenarios may not be relevant for the situation to be assessed within a certain framework. For plant protection products and veterinary pharmaceuticals, for example, the application regime determines the predicted exposure pattern. If the exposure is predicted to be a single peak that declines quickly, a chronic fish study may be less relevant. However, if a substance is present over a longer period of time, because there is continuous discharge into aquatic systems and/or the substance disappears slowly from the water phase, then a chronic fish study may be very relevant.

13. Is the tested exposure scenario relevant for the species? Depending on the framework and the purpose of assessment, the exposure scenario may not be relevant for the species tested. For example, exposure for only a few minutes can be relevant to study reproductive effects in fish eggs but may not be relevant to assess acute or chronic effects on adult fish.

CRED REPORTING RECOMMENDATIONS

The CRED reporting recommendations contain 50 specific criteria divided into 6 categories: general information, test design, test substance, test organism, exposure conditions, and statistical design and biological response (Table 5). The reporting recommendations have been developed to match the reliability criteria from the CRED evaluation method, and the guidance accompanying the evaluation method is therefore also useful for researchers. Researchers performing aquatic ecotoxicity studies are advised to go through the reporting recommendations at an early stage of designing their experiments to make sure that all aspects connected to reliability are considered. Some of the recommendations are critical for the reliability of a particular study; others will be of less importance. Often, this will depend on test organism, test duration, and test substance; for example, it is not relevant to

Table 5. The CRED reporting recommendations contain 50 specific aspects to consider when reporting aquatic ecotoxicity studies[a]

CRED reporting recommendations
1. General information
a. Purpose of study
b. Description of endpoints
2. Test design
a. Performed according to standard/modified standard (e.g., OECD, USEPA)
b. Performed according to good laboratory practices (GLP)
c. Description of control(s): Negative control, solvent control, positive control
d. Control(s) mortality, growth, morbidity, and other observed non-standard effects such as behavior and coloring
e. Comparison to validity criteria (e.g., control survival, growth) from appropriate guideline test method
3. Test compound
a. Identification (e.g., name, CAS number, specify if the salt or the base is tested)
b. Physicochemical characteristics that may influence the behavior of the compound during the study (e.g., solubility, volatility, stability [hydrolysis, photolysis, degradation], solubility, log K_{OW}, degradability, adsorption)
c. Source
d. Purity percentage
e. Composition of product formulation and presence of impurities
4. Test organism
a. Scientific name
b. Body weight, length
c. Age/life stage
d. Growth/reproductive condition
e. Sex
f. Strain, clone
g. Source, including possible pre-exposure for field-collected organisms
h. Culture handling and acclimation to exposure conditions
5. Exposure conditions
a. Exposure schedule (static, semistatic, flow-through system, other) and flow rate (flow-through systems) or renewal time (semistatic systems)
b. Open or closed system
c. Test medium composition and source of test water (e.g., well water, deionized water, tap water)
d. Temperature and time points for measuring
e. pH and time points for measuring
f. Hardness of water and time points for measuring
g. Conductivity and time points for measuring
h. Dissolved oxygen content and time points for measuring
i. Light intensity and quality (e.g., source, light spectrum, homogeneity), light/dark conditions
j. Feeding protocols, food composition
k. Material and volume of aquarium/container and other equipment in contact with test organisms and test substance
l. Use of sand or sediment and its characteristics (total organic carbon, particle size, etc.)
m. Preparation of stock solutions, including solvent concentrations in test water and controls
n. Nominal concentrations of test substance
o. Measured concentrations of test substance and time points for measuring
p. Analytical method: description of method, including limit of detection and limit of quantification
q. Exposure duration and total test duration
r. Time points of observations for endpoints
s. Results based on nominal or measured concentrations
t. Biomass loading (biomass per liter)
6. Statistical design and biological response
a. Number of replicates for control(s) and test concentrations; setup of replicates (avoid pseudoreplication)
b. Number of organisms, or algal cell concentration, per replicate
c. Treatment design (e.g., block, randomized)
d. Statistical method used
e. Biological response for each concentration
f. Dose–response observed
g. Statistically significant responses noted (e.g., ECx)
h. Significance level for NOEC and LOEC data (0.05 or less)
i. Estimation of variability for LCx and ECx data
j. Availability of raw data: through supporting information, a website, or upon request

[a]Authors are encouraged to provide a rationale whenever a reporting recommendation is not met. The CRED reporting recommendations are also available in an Excel file (see Supplemental Data, Appendix A).

CRED = criteria for reporting and evaluating ecotoxicity data; OECD = Organisation for Economic Co-operation and Development; USEPA = US Environmental Protection Agency; K_{OW} = octanol–water partition coefficient; ECx = x% effect concentration; NOEC = no-observed-effect concentraiton; LOEC = lowest-observed-effect concentration; LCx = x% lethal concentration.

report sex of the organism for an algae test, feeding protocols are not relevant for acute ecotoxicity studies, and the age and sex of test organisms are essential when testing endocrine-disrupting substances. When reporting ecotoxicity studies, authors are encouraged to include as much information as reasonably possible in a structured manner, using the supplemental data if necessary. When no information can be provided for one or several of the reporting recommendations, it is suggested that authors transparently explain why the information was not reported. This way, anyone evaluating the study (e.g., peer reviewer, editor, fellow researcher, risk assessor) can get a clear picture of experimental design, results, and possible limitations of a particular study. The possibility that a study is under-reported and essential information is missing is likely to

decrease if the CRED reporting recommendations are followed. In addition, a study containing all important information will probably go through the peer-review process in a more efficient way.

DISCUSSION

CRED evaluation method

The CRED evaluation method is an adaptation from the Klimisch et al. method [3] and distinguishes itself mainly through the use of more detailed criteria in combination with a clear explanation of these criteria, and by providing criteria for evaluation of both reliability and relevance. Within the CRED project, a ring test was performed in which the CRED evaluation method was compared with the Klimisch method [21]. The ring test showed that the reliability categorizations using both methods did not differ significantly for 5 of the 8 studies assessed. The significant differences between the methods for the other 3 studies could be attributed to a more in-depth assessment using the CRED evaluation method; that is, flaws in the study setup and performance were detected more frequently. Fine-tuning of the criteria was mainly focused on rewriting those evaluation criteria for which the answers of the participants differed the most. The CRED evaluation method was also perceived to be more accurate, consistent, and applicable (practical) for routine use compared with the Klimisch method. Moreover, ring test participants found the CRED evaluation method to depend less on expert judgment. An important aspect when implementing a new system is the time needed for the evaluation. The results of the ring test showed that, on average, participants did not need more time using the CRED evaluation method than when using the Klimisch et al. method (see Supplemental Data), even though most of the participants were already familiar with the Klimisch et al. method and not with the CRED evaluation method.

As a result, 80% of the participants felt "very confident" or "confident" when using the CRED evaluation method for reliability evaluation, in comparison with 60% using the Klimisch method. For relevance evaluations, 72% felt "very confident" or "confident" when using the CRED evaluation method, compared with 37% using the Klimisch method (see also Kase et al. [21] for more details). Thus, the CRED evaluation method is a good replacement of the Klimisch method. Since the reliability categories are the same (reliable without restrictions, reliable with restrictions, not reliable, not assignable), former evaluations using the Klimisch method can still be used and do not necessarily need to be replaced by an evaluation using the CRED method. However, use of the CRED evaluation method, with its increased transparency, will facilitate the exchange of assessments among frameworks.

It should be stressed that, despite the detailed guidance given in the present study, use of expert judgment while evaluating studies is still necessary. Determining reliability is not a box-checking exercise, where the number of passed or failed criteria is determined to obtain a reliability category [32]. Any method used to assess the reliability and relevance of a study should be based on sound scientific argumentation, and expert judgment is essential.

During the ring test, the CRED criteria for reliability and relevance were categorized as either critical or noncritical. However, this appeared to create confusion among the ring test participants since many of them applied these criteria in a very strict sense, despite the accompanying text explaining possible exceptions. In most cases, whether a criterion is critical or not is not a black and white decision but depends on the substance and/or species tested. For example, a closed system is essential for volatile substances, the light conditions are essential for algae, and it is essential that concentrations are measured and a flow-through or renewal system is used for some unstable substances, while a static system with nominal concentrations may be good enough for stable substances. Thus, the division into critical and noncritical criteria is no longer part of the CRED evaluation method. Although the CRED evaluation method is more detailed than the Klimisch method, it still does not give a solution for every possible situation. For most of the criteria, expert judgment is needed to decide whether a certain criterion is critical for the specific study under evaluation.

The CRED ring test provides an illustrative example of how the same ecotoxicity study can be assessed in various ways. Ring test participants were asked to evaluate 2 studies that relied on the same raw data: a contract laboratory study report and a peer-reviewed publication [21]. When the laboratory study report was evaluated using the Klimisch method ($n = 9$), it was assessed to be either "reliable without restrictions" (R1; 44% of the participants) or "reliable with restrictions" (R2; 56% of the participants), meaning that it could have been used for regulatory purposes according to all ring test participants. However, when the same study was evaluated using the more detailed CRED evaluation method ($n = 19$), participants had a more critical look and identified flaws in the study, which resulted in a reduced reliability: 16% of the participants evaluated the study to be "reliable without restriction" (R1), 21% evaluated the study to be "reliable with restrictions" (R2), and 63% came to the conclusion that the study was "not reliable" (R3). The difference in assignment of categories was found to be statistically significant using a chi-squared test ($p = 0.005$). A possible explanation for this could be that risk assessors are biased toward laboratory studies performed according to guideline studies and GLP. Also, when using the Klimisch method, participants rated the contract laboratory study report much higher ($n = 9$; 44% R1 and 56% R2) than the peer-reviewed publication based on the same data set ($n = 12$; 66% R2 and 33% R3). When the CRED evaluation method was used, the contract laboratory study report was overall assigned a lower reliability category ($n = 19$; 16% R1, 21% R2, 63% R3) than when the Klimisch method was used, but the difference in assignment of categories for the peer-reviewed study using CRED ($n = 13$; 46% R2, 31% R3, 23% R4) was not statistically significant ($p = 0.321$) compared with the Klimisch evaluation. Thus, a critical view combined with the CRED evaluation method will prevent an automatic assignment of "not reliable" (R3) to peer-reviewed publications and "reliable without restriction" (R1) to contract laboratory study reports. The CRED evaluation method uses the same reliability and relevance criteria for guideline studies and peer-reviewed studies.

In addition to emphasizing the problems with bias among risk assessors and the importance of having a sufficiently detailed evaluation method, this example also shows that poorly performed studies and insufficient reporting are not limited to the work conducted within the peer-reviewed literature but also can be found in industry-generated guideline studies performed according to GLP. In addition, assessors should be aware that laboratory study reports may still be unreliable even if they provide all necessary information to evaluate reliability (e.g., if the test design does not take the physicochemical properties into account, analytical verification is poor, or solvent controls did not perform well).

The CRED evaluation method aids in making the assessors' decisions more consistent and, moreover, helps with documenting the choices made and increases transparency in study evaluations. However, as stated before, use of the CRED evaluation method should not be just a box-checking exercise. When performing evaluations, some flexibility may be required, especially for older studies since reporting of studies has not always been done with regulatory use in mind. If the CRED evaluation method is used in a too strict manner, much data may get lost and the consequences of this may be severe for data-poor substances. Although this might be inevitable in the short run, we hope that with the CRED evaluation method and the CRED reporting recommendations the reliability of peer-reviewed articles will improve and nearly all of them can be available for risk assessment or EQS derivations in the future.

When using the CRED evaluation method, the transparency of study evaluations increases, which has the potential to also increase the exchange of assessments between frameworks if the risk assessors' evaluations are documented in a structured manner. To facilitate the documentation, an Excel document is available as Supplemental Data, Appendix B. To aid transparency, the reason for fulfillment or nonfulfillment should be stated for each criterion. The improved guidance in combination with a more transparent evaluation system also prevents the often-made mistake of assigning a study to be "not reliable" (R3) when the only problem is that it lacks sufficient details. This was shown in the ring test, where a study that lacked the details for a thorough assessment was frequently assigned to be "not reliable" (R3) using the Klimisch method but more frequently assigned to be "not assignable" (R4) using the CRED evaluation method.

Finally, we stress the following general recommendations for risk assessors: evaluate reliability and relevance in a systematic way, such as by using the checklist provided in the Supplemental Data; document your choices and the rationale behind them, highlight uncertainties, and stay critical regarding bias coming from the background of the authors; contact the authors if more details are needed; derive other effect values (e.g., EC10) when data are available; make sure all derivations are consistent; use an expert group for review if possible; and let an agency from another country review your derivation.

CRED reporting recommendations

Peer-reviewed ecotoxicity studies and contract laboratory study reports have a similar purpose: to provide the reader with information about the test and the results. However, the content and structure of the reports differ. Guided by limited publication space and word limits, peer-reviewed literature often lacks the necessary details for a thorough evaluation of the study [4,15].

There are several aspects that affect scientific credibility: source of funding, conflicts of interests, choice of scientific method, interpretation of results, unbiased reporting, use of peer review, and public access to raw data [33–36]. A key issue for all these aspects is transparency. Transparent reporting of ecotoxicity studies can be achieved when details about test design and results are reported in a structured way using reporting recommendations. The CRED reporting recommendations safeguard that at least a minimum amount of information regarding the study design and study results is reported. In addition, the recommendations could simplify the writing process by serving as a template for authors. Furthermore, by following the CRED reporting recommendations when designing experiments, researchers can safeguard the reliability of ecotoxicity studies. Improved reporting of ecotoxicity studies could also facilitate the possibility of other researchers reproducing test results [19,37,38], thus increasing the credibility of the scientific process as a whole. It also has the potential to accelerate the publication process by streamlining editors' and reviewers' work. Structured reporting recommendations can also create awareness regarding the requirements set by regulatory agencies for inclusion of studies in risk assessments [23] and can simplify the evaluation process in regulatory assessment of chemicals. Finally, researchers should keep in mind that nonguideline studies, which may or may not be performed according to GLP, also can be used for regulatory purposes [26,27]. A concise and complete description of methods and results (if necessary in the supplemental data) is essential for the publication process but also for the study to be used by others and for other purposes.

Supplemental Data—The Supplemental Data are available on the Wiley Online Library at DOI: 10.1002/etc.3259.

Acknowledgment—We thank T. Vermeire, J. Ferreira, and E. Smit (RIVM) as well as I. Werner (Swiss Centre for Applied Ecotoxicology EAWAG-EPFL) and the anonymous reviewers for valuable comments during the preparation of the manuscript. M. Ågerstrand's research is funded by Mistra (Swedish Foundation for Strategic Environmental Research).

Conflict of interest—The authors declare no conflicts of interests.

Data availability—Data are available on request to the authors (caroline.moermond@rivm.nl).

REFERENCES

1. Beronius A, Rudén C, Håkansson H, Hanberg A. 2010. Risk to all or none? A comparative analysis of controversies in the health risk assessment of bisphenol A. *Reprod Toxicol* 29:132–146.
2. Rudén C. 2002. From data to decision. A case study of comtroversies in cancer risk assessments. PhD thesis. Karolinska Institutet, Solna, Sweden.
3. Klimisch H-J, Andreae M, Tillmann U. 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regul Toxicol Pharmacol* 25:1–5.
4. Ågerstrand M, Breitholtz M, Rudén C. 2011. Comparison of four different methods for reliability evaluation of ecotoxicity data: A case study of non-standard test data used in environmental risk assessments of pharmaceutical substances. *Environ Sci Eur* 23:17.
5. Tweedale T, Lysimachou A, Muilerman H. 2014. Missed and dismissed: Pesticide regulators ignore the legal obligation to use independent science for deriving safe exposure levels. PAN Europe, Brussels, Belgium.
6. Hobbs DA, Warne MSJ, Markich SJ. 2005. Evaluation of criteria used to assess the quality of aquatic toxicity data. *Integr Environ Assess Manage* 1:174–180.
7. Durda JL, Preziosi DV. 2000. Data quality evaluation of toxicological studies used to derive ecotoxicological benchmarks. *Hum Ecol Risk Assess* 6:747–765.
8. US Environmental Protection Agency. 2011. Evaluation guidelines for ecological toxicity data in the open literature. [cited 2015 June 2]. Available from: http://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/evaluation-guidelines-ecological-toxicity-data-open
9. European Commission. 2005. The appropriateness of existing methodologies to assess the potential risks associated with engineered and adventitious products of nanotechnologies. SCENIHR/002/05. Brussels, Belgium.
10. Handy RD, Owen R, Valsami-Jones E. 2008. The ecotoxicology of nanoparticles and nanomaterials: Current status, knowledge gaps, challenges, and future needs. *Ecotoxicology* 17:315–325.
11. Swedish Medical Product Agency. 2004. Miljöpåverkan från läkemedel samt kosmetiska och hygieniska produkter. [cited 2015 June 2]. Available from: https://lakemedelsverket.se/upload/om-lakemedelsverket/publikationer/040824_miljouppdraget-rapport.pdf
12. Allanou R, Hansen BG, van der Bilt Y. 2003. Public availability of data on EU high production volume chemicals. Part 1. [cited 2013 May 27]. Available from: http://publications.jrc.ec.europa.eu/repository/handle/111111111/1075

13. Environmental Defense Fund. 1997. Toxic ignorance: The continuing absence of basic health testing for top-selling chemicals in the United States. Washington, DC.
14. European Chemicals Agency. 2011. The use of alternative to testing on animals for the REACH regulation. ECHA-14-A-07-EN. Helsinki, Finland.
15. Ågerstrand M, Edvardsson L, Rudén C. 2013. Bad reporting or bad science? Systematic data evaluation as a means to improve the use of peer-reviewed studies in risk assessments of chemicals. *Hum Ecol Risk Assess* 20:1427–1445.
16. Miller GW. 2014. Improving reproducibility in toxicology. *Toxicol Sci* 139:001–003.
17. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M; for the STROBE Initiative. 2007. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. *PLoS Med* 4:e297.
18. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. 2010. Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. *PLoS Biol* 8:e1000412.
19. *Nature*. 2013. Editorial—Announcement: Reducing our irreproducibility. *Nature* 496:398–398.
20. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. 2001. Minimum information about a microarray experiment (MIAME)—Toward standards for microarray data. *Nat Genet* 29:365–371.
21. Kase R, Korkaric M, Werner M, Agerstrand M. 2016. Criteria for Reporting and Evaluating ecotoxicity Data (CRED): Comparison and perception of the Klimisch and CRED methods for evaluating reliability and relevance of ecotoxicity studies. *Env Sci Eur* 28:1–14.
22. Schneider K, Schwarz M, Burkholder I, Kopp-Schneider A, Edler L, Kinsner-Ovaskainen A, Hartung T, Hoffmann S. 2009. "ToxRTool", a new tool to assess the reliability of toxicological data. *Toxicol Lett* 189:138–144.
23. Ågerstrand M, Küster A, Bachmann J, Breitholtz M, Ebert I, Rechenberg B, Rudén C. 2011. Reporting and evaluation criteria as means towards a transparent use of ecotoxicity data for environmental risk assessment of pharmaceuticals. *Environ Pollut* 159:2487–2492.
24. Mensink BJWG, Smith CE, Montforts MHMM. 2008. Manual for summarizing and evaluating aspects of plant protection products. Report number 601712006/2010. RIVM, Bilthoven, The Netherlands.
25. European Chemicals Agency. 2012. *How to Report Robust Study Summaries: Practical Guide 3*, Version 2.0. Helsinki, Finland.
26. European Chemicals Agency. 2008. REACH guidance documents. Helsinki, Finland.
27. European Commission. 2011. Common implementation strategy for the Water Framework Directive (2000/60/EC). Guidance document No. 27. Technical guidance for deriving environmental quality standards. [cited 2015 June 2]. Available from: https://circabc.europa.eu/sd/a/0cc3581b-5f65-4b6f-91c6-433a1e947838/TGD-environmental quality standard%20CIS-WFD%2027%20EC%202011.pdf
28. European Commission. 2002. Guidance document on aquatic ecotoxicology in the context of the Directive 91/414/EEC. Working Document. Brussels, Belgium.
29. Blanck H, Wängberg S-Å, Molander S. 1988. Pollution-induced community tolerance (PICT)—A new ecotoxicological tool. In Cairns J Jr, Pratt JR, eds, *Functional Testing of Aquatic Biota for Estimating Hazards of Chemicals*. ASTM STP 988. American Society for Testing and Materials, Philadelphia, PA, pp 219–230. [cited 2015 June 3] Available from: http://publications.lib.chalmers.se/publication/139772-pollution-induced-community-tolerance-pict-a-new-ecotoxicological-tool-in-cairns-j-jr-pratt-jr-eds-f
30. Organisation for Economic Co-operation and Development. 2012. Information on OECD work related to endocrine disrupters. Paris, France. [cited 2015 June 2]. Available from: http://www.oecd.org/env/ehs/testing/50067203.pdf
31. Hutchinson TH, Ankley GT, Segner H, Tyler CR. 2006. Screening and testing for endocrine disruption in fish—Biomarkers as "signposts," not "traffic lights," in risk assessment. *Environ Health Perspect* 114(Suppl. 1): 106–114.
32. Baker M. 2015. US societies push back against NIH reproducibility guidelines. *Nature News*, April 17, 2015. DOI: 10.1038/nature.2015.17354.
33. Harris CA, Scott AP, Johnson AC, Panter GH, Sheahan D, Roberts M, Sumpter JP. 2014. Principles of sound ecotoxicology. *Environ Sci Technol* 48:3100–3111.
34. Grandjean P. 2013. Problems with hidden COI. *Scientist* [cited 2014 February 26]. Available from: http://www.the-scientist.com/?articles.view/articleNo/37934/title/Opinion--Problems-with-Hidden-COI/.
35. Conrad JW Jr, Becker RA. 2011. Enhancing credibility of chemical safety studies: Emerging consensus on key assessment criteria. *Environ Health Perspect* 119:757–764.
36. Bohannon J. 2013. Who's afraid of peer review? *Science* 342:60–65.
37. Ioannidis JPA, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, Schulz KF, Tibshirani R. 2014. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* 383:166–175.
38. Tilson HA, Schroeder JC. 2013. Reporting of results from animal studies. *Environ Health Perspect* 121:A320–A321.