DTU Library

NanoCRED: A transparent framework to assess the regulatory adequacy of ecotoxicity data for nanomaterials – relevance and reliability revisited

Hartmann, Nanna B.; Ågerstrand, Malene; Lützhøft, Hans-Christian Holten; Baun, Anders

Link back to DTU Orbit

NanoCRED: A transparent framework to assess the regulatory adequacy of ecotoxicity data for nanomaterials – relevance and reliability revisited

Nanna B Hartmann†*, Marlene Ågerstrand‡, Hans-Christian Holten Lützhøft†, A Baun†

†Department of Environmental Engineering, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark (nibh@env.dtu.dk)

‡ Department of Environmental Science and Analytical Chemistry, Stockholm University, SE-106 91 Stockholm, Sweden

*To whom correspondence may be addressed. E-mail: nibh@env.dtu.dk Telephone: +45-45251401; Fax: +31-30-2744401

ABSTRACT

Environmental hazard and risk assessment serve as the basis for regulatory decisions to protect the environment from unintentional adverse effects of chemical substances including nanomaterials. This process requires reliable and relevant environmental hazard data upon which classification and labelling can be based and Predicted No-Effect Concentration (PNEC) values can be estimated. In a regulatory context ecotoxicological data is often recommended to be generated according to accepted and validated test guidelines, preferably also following Good Laboratory Practice. However, engineered nanomaterials are known to behave very differently in ecotoxicity tests compared to the conventional soluble chemicals, for which most guidelines were developed. Therefore non-guideline tests, or tests following modified test guidelines, can provide valuable information and should not *per se* be considered less adequate for regulatory use. Here we propose a framework for reliability and relevance evaluation of ecotoxicity data for nanomaterials that take into account the challenges and characterisation requirements associated with testing of these substances. The nanoCRED evaluation criteria, and accompanying guidance, were developed to be used in combination with those developed through the 'Criteria for Reporting and Evaluating Ecotoxicity Data (CRED)' project. This approach can accommodate all types of nanomaterials, all types of aquatic ecotoxicity studies, and qualitative as well as quantitative data evaluation requirements. Furthermore, it is practically feasible to implement and directly applicable in European as well as international regulatory frameworks.

Key words – nanoecotoxicology, chemical risk assessment, data evaluation, REACH, Predicted No-effect Concentration

1. INTRODUCTION

1.1. Environmental hazard and risk assessment of nanomaterials

The increasing use and anticipated environmental release of engineered nanomaterials[1] from industrial and consumer applications makes environmental hazard and risk assessment of these substances of highest importance. Environmental risk assessment is a legal requirement in many countries to ensure safe production and use of chemicals and involves the gathering and critical evaluation of ecotoxicological data. For example, the United States Environmental Protection Agency (US-EPA) requires a pre-manufacture notice prior to the production or import of new chemical substances, including nanomaterials, under the Toxic Substances Control Act. Based on the submitted information, it is assessed if the chemical substance will pose a risk to human health or the environment and/or if further assessment is required (USEPA, 2016). In the European Union the procedure for chemical safety assessment is comprised of an information gathering followed by I) an exposure assessment for deriving a Predicted Environmental Concentration (PEC) and II) a dose-response (effects) assessment leading to classification and labelling (EC, 2008; UN, 2011), identification of so-called PBT (persistent, bioacummulative and toxic) compounds (ECHA , 2014; ECHA, 2015; ECHA 2016; UN, 2011) and determination of a Predicted No-Effect Concentration (PNEC).

These procedures have been developed for conventional (soluble) chemicals, but are also considered principally applicable to nanomaterials (Aitken et al., 2011). It is well-known, however, that ecotoxicity testing of nanomaterials is not straight-forward since they behave very differently in ecotoxicity test systems compared to the conventional soluble chemicals, for which the existing test guidelines and guidance documents were intended (Pettitt et al., 2013). Their applicability to

---

[1] With 'nanomaterials' often defined as materials having one or more dimensions in the size range of 1-100 nm (e.g. EU, 2011).

nanomaterial testing can therefore be questioned (Khan et al., 2017; Kühnel & Nickel 2014; Hund-Rinke et al., 2016). The fundamental reason is that nanomaterials are often insoluble or only partly soluble in water. Although they may release dissolved chemical species, they remain present as a solid phase forming a suspension of particles that may undergo physical transformation processes such as agglomeration and aggregation. Release of dissolved ions is well-known for nanoparticles such as ZnO, CuO, and Ag for which the contribution of the dissolved species and/or the particulate form of the metals remain a topic of discussion (Skjolding et al. 2016; Notter et al. 2014; Brinch et al. 2016).

Nanomaterials thus have properties and behaviours in aquatic test systems that are different from conventional chemicals. Transformation processes, which occur before and during ecotoxicity testing, will change the exposure conditions, often violating the assumption of stable exposure concentrations during testing. It can be argued that ecotoxicity data, produced using current guideline tests, are not sufficiently sensitive (Skjolding et al., 2016) and that many of the currently established test guidelines need to be adapted or re-designed to be applicable to nanomaterials (Hund-Rinke et al. 2016; Kühnel & Nickel 2014). For example, commonly prescribed test organisms may not be ideal for the detection of non-chemical effects of nanomaterials (Khan et al., 2017). The OECD is a key player in this development and, although current OECD Test Guidelines for ecotoxicity are in principal applicable to nanomaterials, guidance on nano-specific testing issues is currently lacking (although under development) (Petersen et al., 2015; Hansen et al., 2017). Until these are published, regulatory assessments will continue to rely on studies performed according to existing test guidelines and lessons learned from the open scientific literature. In the meantime - and also enabling future use of existing non-guideline ecotoxicity data - a systematic and transparent method to evaluate data adequacy is needed to support data selection for environmental risk assessment of nanomaterials.

1.2. Current methods for evaluation of ecotoxicity studies for regulatory use

When evaluating the regulatory adequacy of ecotoxicity studies for hazard and risk assessment purposes two elements are included in the assessment: reliability and relevance. Reliability is defined as "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" (ECHA, 2008). Reliability is hence determined by a combined assessment of test performance, data analysis and study documentation, and that all three parameters have to be sufficiently fulfilled for a study to be considered reliable. This implies that a seemingly scientifically sound study may be considered not assignable if it is not sufficiently documented. In the same way a well-documented study is considered less reliable if the study design or test performance is inadequate or considered not to be appropriate for the tested substance. In addition, the reliability of an ecotoxicity study does not change depending on the type of hazard and risk assessment for which the study is used (Moermond et al., 2016).

Relevance is defined as "covering the extent to which data and tests are appropriate for a particular hazard identification or risk characterization" (ECHA, 2008), meaning that the relevance of an ecotoxicity study can vary. A study can for example be very relevant for an aquatic risk assessment, but less relevant for a soil risk assessment. The relevance is determined by an evaluation of e.g. the chosen endpoint, test organism, and test design in relation to the regulatory purpose of the assessment.

In the REACH guidance for chemical safety assessment it is recommended that the reliability of data is assessed carefully, and that this could be done using the evaluation method described by Klimisch et al. (1997). The Klimisch method offers twelve reliability evaluation criteria for acute studies and fourteen for chronic studies. However, no additional guidance on these criteria is included, nor does it specify criteria for evaluation of relevance. A reliability evaluation according to the Klimisch method results in one of the following reliability categories: 'Reliable without

restrictions', 'Reliable with restrictions', 'Not reliable', and 'Not assignable'. In most regulatory risk assessment frameworks, studies assigned reliable with or without restrictions are considered acceptable (Küster et al. 2009). The Klimisch method has been criticized for not being detailed enough to qualify as a sufficient tool for evaluators, and for favouring standard test methods and GLP-studies over non-standard studies from the peer-reviewed literature (Ågerstrand et al. 2011; Kase et al. 2016).

Several alternative methods for evaluating the reliability and relevance of (eco)toxicity studies have been described in peer-reviewed literature (Hobbs et al. 2005; Schneider et al. 2009; Durda & Preziosi 2000; Ågerstrand et al. 2011; Küster et al. 2009). A study by Ågerstrand et al. (2011) compared four of these methods and found, not surprisingly, that the choice of evaluation method affects the outcome of a study evaluation. Evaluation methods that lack in detail will depend more on expert judgment, which is *per se* not a disadvantage for the evaluation but could result in biased evaluations for example when certain aspects are prioritized differently depending on the evaluator's expertise. A structured and detailed evaluation method with well-defined criteria ensures that a similar set of criteria are considered in each evaluation, which improves consistency of evaluations, i.e. so that several evaluators arrive at the same, or at least similar, conclusion when evaluating the same study. A structured and detailed evaluation method can also improve transparency of evaluations, thereby enabling a thorough understanding of how decisions were made for a third-party reviewer (Ågerstrand et al. 2013; Kase et al. 2016). As a response to the growing critique against the Klimisch method and general inconsistencies in data evaluation methods, the CRED (Criteria for Reporting and Evaluating ecotoxicity Data) evaluation method was developed (Moermond et al. 2016)

1.3. Evaluation of nanomaterial ecotoxicity studies

Neither the Klimisch method nor the CRED evaluation method offer nano-specific guidance for evaluation of ecotoxicity studies. The challenges of evaluating the reliability of ecotoxicological data for nanomaterials are also reflected in the information provided in REACH registration dossiers for nanomaterials, according to an assessment by the European Commission through the NANO SUPPORT Project (EC, 2012). The project assessed nanomaterial dossiers that were submitted by the first REACH registration deadline in March 2011. The Final Report of the project highlights a lack of transparency and justifications in the choice of 'key studies' and 'supporting studies for the different endpoints and also describes an inconsistency in how reliability categories were appointed to different studies. For example, in one of the assessed REACH dossier, an algal growth inhibition test (performed according to the OECD 201 guideline and ISO 6341 standard) was considered 'not reliable' as the study setup was unsuitable for testing of nanomaterial effects. However, in another dossier this study setup was considered to be 'reliable without restrictions' (EC, 2012).

For human health risk assessments there are examples of evaluation methods to assess data quality for nanomaterial studies, e.g. a two-step process for evaluation of nanotoxicity studies developed by Card and Magnuson (2010). In their procedure the first step is a reliability evaluation performed using the ToxRTool (Toxicological data Reliability Assessment Tool) developed by Schneider et al. (2009). This is done with the purpose of providing "comprehensive criteria and guidance for evaluations of the inherent quality of toxicological data, thus making the decision process of assigning reliability categories more transparent and more harmonised" (JRC, 2017), i.e. assignment of a Klimisch category based on a transparent evaluation scheme. In the second step, a nano-score is calculated by evaluating the documentation of nano-specific characteristics such as agglomeration, surface charge and particle size. The results from the two steps are then combined into a final nano-score for the reliability of the evaluated study.

This method by Card and Magnuson (2010) is directed towards human health toxicity data. Other schemes for evaluating and scoring nanomaterial data have recently been reviewed (Robinson et al., 2016) highlighting e.g. the 'Literature Criteria Checklist' developed by the DaNa project (DaNa, 2016) and a scheme to assess the data quality for nanotoxicity studies to be used for QSAR modelling developed by Lubinski et al. (2013). So far, however, no method for evaluation of the regulatory adequacy of ecotoxicity data for nanomaterials exists. The aim of this paper was therefore to develop NanoCRED, a transparent and structured framework for evaluation of the reliability and relevance of nanomaterial ecotoxicity data. The intention is to provide a science based nano-specific evaluation criteria and guidance material for regulatory adequacy of aquatic (pelagic) ecotoxicity data for nanomaterial. The method aims to be on the one hand practical and flexible enough to accommodate all types of nanomaterials and all types of aquatic ecotoxicity studies, and on the other hand directly applicable to data selection for regulatory purposes.

2. METHODS

The evaluation method for nanomaterial ecotoxicity studies (i.e. nanoCRED) was developed largely on the basis of the method developed in the CRED project (Moermond et al. 2016), drawing on nano-specific test recommendations specified by OECD (2012), criteria included in existing evaluation tools for human health toxicity data for nanomaterials (such as those proposed by the Card and Magnusson, 2010), criteria established by the DaNa project (DaNa, 2016), and information from the open scientific literature. When adjusting and refining the criteria for evaluation of aquatic nanomaterial ecotoxicity data additional needs for data on physico-chemical properties as well as the characterization before and during testing (OECD, 2012) were taken into account. In this method we adjusted the CRED reliability categories, as well as a subset of the CRED reliability criteria and the guidance material. In addition, for certain relevance criteria additional nano-specific guidance was included.

3. RESULTS

3.1. Reliability evaluation for nanomaterials

The reliability categorisation of nanomaterial ecotoxicity studies is based on a) the appropriateness of the study design for the purpose of nanomaterial testing, b) documentation provided on the study design, c) data on inherent nanomaterial properties and d) nanomaterial characterization in the test system and during the test. Descriptions of the nanoCRED categories (nR1-nR4) are presented in Table 1.

Table 1. Nano-specific reliability categories (adapted and further developed from Klimisch et al. (1997) and Moermond et al. (2016)). For information on critical (▲▲▲) and important (▲▲) reliability criteria see Table 2.

Score	Description
nR1	**Reliable without restrictions**: All critical (▲▲▲) and important (▲▲) reliability criteria are fulfilled or partially fulfilled. The study is well designed, performed and documented. Nanomaterial properties and behaviour in the test system is extensively documented. The experiment has been carried out according to methods that are considered scientifically appropriate for ecotoxicity testing of nanomaterials and where the physico-chemical properties of the nanomaterial are considered in the test design. If (when) specific nanomaterial guidance or guidelines exist, the use of these may be considered favourable.
nR2	**Reliable with restrictions**: Most critical (▲▲▲) and important (▲▲) criteria are fulfilled or partially fulfilled. The study is generally well designed, performed and documented, but some minor flaws in the documentation or setup may be present. Nanomaterial properties and behaviour in the test system is well documented. The experimental design and test method are considered scientifically appropriate for ecotoxicity testing of nanomaterials but may contain some minor flaws in documentation or setup.
nR3	**Not reliable:** Not all critical reliability criteria are fulfilled or partially fulfilled. This mainly concerns studies which have clear flaws in study design and study conduction, and/or where the experimental design and test method are considered not to be scientifically appropriate for ecotoxicity testing of nanomaterials.
nR4	**Not assignable:** Information needed to make an assessment of one or more critical and important criteria is missing. This concerns studies or data from the literature which do not give sufficient experimental details, or reports where the documentation is not sufficient for assessment of reliability for one or more critical parameters.

Nano-specific considerations are provided for criterion number 1-7, 10, 12-13, 15-16, and 19 of the CRED criteria. Criterion number 8, 9, 11, 14, 17, 18 and 20 do not differ from the general CRED criteria. For guidance on these criteria see Moermond et al. 2016 and Supporting Information. All criteria are presented in Table 2. In contrast to the CRED evaluation method, the reliability criteria for nanomaterial ecotoxicity studies have been weighted as 'critical' (▲▲▲), 'important' (▲▲), or 'of minor importance' (▲) to guide the evaluator and increase the applicability of the method (Table 2). Rationale behind the weighing is provided in the specific guidance text for each criterion. It should be kept in mind that the weighing is based on expert judgement and the current scientific understanding of nanomaterial behaviour and effects in ecotoxicity tests. The weighing may therefore be changed as a result of new information. It is therefore suggested that the evaluator scrutinize the reliability criteria and evaluate if any weighing should be adjusted due to newly emerged information on ecotoxicity testing of nanomaterials. There may also be case-specific reasons to make such changes, or, for some criteria, a possibility to choose between different pre-defined options. In these cases, such changes or choices should be justified and documented. The suggested weighing is applicable only to studies on nanomaterials, and should therefore not be used in the evaluation of tests with soluble substances.

Table 2. *Reliability criteria (adopted from Moermond et al., 2016). Nano-specific guidance is available in the main text for criteria 1-7, 10, 12-13, 15-16 and 19. For general explanatory guidance text on how to interpret these criteria see SI.* **Criticality level.** *▲ ▲ ▲ : These criteria are critical for study reliability, ▲ ▲ : These criteria are important for study reliability, ▲ : These criteria are of minor importance for study reliability, but may support study evaluation, (▲/▲ ▲): Importance of these criteria depend on the specific test and/or nanomaterial properties.*

Number	Criticality level (▲ ▲ ▲/▲ ▲/▲)	Criterion
		General information
		Before evaluating the test, check the physico-chemical characteristics of your nanomaterial (handbooks/general sources/scientific litterature).
		Test setup
1	▲	Is a guideline method (e.g., OECD/ISO) or modified guideline used?
2	▲	Is the test performed under GLP conditions?
3	▲/▲ ▲	If applicable, are validity criteria fulfilled (e.g. control survival, growth)?
4	▲ ▲ ▲	Are appropriate controls performed (e.g. dispersant control, metal ion control, larger than nano-sized (bulk) particles of the same chemical composition, negative and positive control)?
		Test Compound
5	▲ ▲ ▲	Is the tested nanomaterial appropriately identified with name or chemical identifier (e.g., CAS-number)? Are nanomaterial characteristics reported that allow for a clear identification of the tested nanomaterial (e.g. particle size, shape, particle size distribution, surface area, coating)? Are test results reported for the relevant nanomaterial?
6	▲ ▲	Is the purity of the tested nanomaterial reported? This includes information on synthesis by-products as well as synthesis catalysts and presence of other crystalline forms of the substance. And/or, is the source of the test substance trustworthy?
7	▲/▲ ▲	If a formulation is used or if impurities or coatings are present: Do other ingredients in the formulation, the impurities or the coatings exert an effect? Is the amount of nanomaterial in the formulation

known?

Test Organism

| 8 | ▲▲ | Are the organisms well described (e.g. scientific name, sex, 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 the nanomaterial or other unintended stressors? |

Exposure Conditions

10	▲▲	Is the experimental system appropriate for the tested nanomaterial, taking into account its physico-chemical characteristics? Were measures taken to minimise losses and keep exposure stable (in quantitative and qualitative terms) throughout the duration of the test?
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	▲/▲▲	If a dispersant/stabiliser/solvent is used, is the dispersant/stabilizer/solvent within the appropriate range and is a dispersant/stabiliser/solvent control included?
13	▲	Is a correct spacing between exposure concentrations applied?
14	▲▲▲	Is the exposure duration defined?
15	▲▲▲	Have analyses been performed to verify exposure, e.g. concentrations and physico-chemical transformations of the nanomaterial over the duration of the test? In case of unstable exposure, has this been accounted for in the data interpretation? Is the nanomaterial dispersion procedure described in detail?
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)?

3.1.1. NanoCRED reliability criteria and guidance

1 - Is a standard method (e.g., OECD, ISO) or modified standard used? (Criticality level: ▲)

As for conventional chemicals the use of standard methods for ecotoxicity testing of nanomaterials is not in itself a guarantee for study reliability. Use of standard methods may ensure better and more complete documentation but does not per se ensure e.g. an appropriate study design or correct interpretation of test results (Moermond et al. 2016). This is especially true for nanomaterials. This is due to the fact that current guidelines were developed mainly for soluble chemicals and do not take into account the particulate nature of nanomaterials. Until standard methods are available explicitly for testing of nanomaterials, the use of standard methods may therefore result in lower reliability compared to non-standard tests – or modified guideline tests. A well-considered modification of a standard method is preferred over an uncritical use of a standard method that does not account for the nano-specific physical and chemical properties. The use of standard guidelines is therefore considered to be of minor importance for reliability per se.

2 - Is the test performed under GLP conditions? (Criticality level: ▲)

As for standard guidelines, GLP helps to ensure reproducibility and transparency but is not in itself a guarantee for reliability. The use of GLP is therefore considered to be a criterion of minor importance for study reliability.

3 - If applicable, are validity criteria fulfilled (e.g. control survival, growth)? (Criticality level: ▲/▲▲)

This criterion especially applies to studies that follow guidelines or modified guidelines containing requirements regarding validity criteria. However, any study with control group mortality may indicate a test set-up problem. For nanomaterials modifications of the guidelines are often needed to accommodate the different behaviour of the nanoparticles compared to conventional soluble chemicals, for which most of the currently available test guidelines were originally intended. For modified guidelines the original validity criteria may not be relevant or possible to achieve. In these

cases, expert judgement is required. For a description of general issues related to this criterion see Moermond et al. (2016).

4 - Are appropriate controls performed (e.g., dispersant control, metal ion control, larger than nano-sized (bulk) particles of the same chemical composition, negative and positive control)? (Criticality level: ▲ ▲ ▲)

For conventional chemicals it is common to use solvents to dissolve poorly soluble compounds. For nanomaterials it is not feasible and/or relevant to dissolve the particles. Instead the aim is rather to achieve a stable and homogeneous suspension (Hartmann et al., 2015). For example, when testing silver nanoparticles, the aim is not to dissolve the particles into ions and ion complexes but rather to test silver in its nano-particulate form. Stable and homogeneous dispersions can sometimes be achieved by adding a dispersant or a stabiliser to the media. This could be a substance that adsorbs to the surface of the particle and causes electrostatic or steric stabilisation. It is important to include dispersant controls to ensure that this substance does not in itself cause toxic effects. If a dispersant has been used, but no information is provided regarding its ecotoxicity, the study is considered to be "not reliable" (nR3).

Some nanomaterials will dissolve partly in aqueous testing media during testing. This is for example the case for some silver and zinc oxide nanoparticles in standard algal and daphnia media (e.g., Cupi et al., 2015; Sørensen & Baun, 2014). Other nanomaterials, like TiO2, will not undergo dissolution under the testing conditions of ecotoxicity test. For nanomaterials which release metal ions it is important that the study also addresses the metal ion contribution to the observed toxicity. This is often done by including metal ion controls and monitoring the ion release from the nanomaterial. The key to data interpretation is then to evaluate the particle as well as the ion contributions to the observed ecotoxicological effects. Expert judgement is therefore needed to evaluate if the study is lacking critical information regarding metal ion controls.

Nanomaterials are known to introduce certain test artefacts leading to potential misinterpretation of results (Selck et al., 2016; Sørensen et al., 2015). For this reason other types of controls, may be relevant to address e.g. effects caused by endotoxins (as a material impurity), indirect effects (such as effects on nutrient or light availability), photoactivity, fluorescence or absorbance properties (influencing assay measurement) (Selck et al., 2016; Petersen et al., 2014). These and other potential control experiments are described in detail in Petersen et al. (2014).

A negative control refers to test samples in pure media without the presence of nanomaterials, dispersants, etc. Mortality in the negative controls may indicate some problems with the study reliability and result in a lower reliability score (nR2 or nR3) based on expert judgement. Positive controls are useful to check the sensitivity of the test organisms and test performance and are often included as validity criteria in standard guideline tests (see criterion 3). If positive controls show an abnormal response this should be taken into account when evaluating the reliability of a study.

5 - Is the tested nanomaterial identified with name or CAS-number? Are nanomaterial characteristics reported that allow for a clear identification of the tested nanomaterial (e.g. particle size, shape, particle size distribution, surface area, coating)? Are test results reported for the relevant nanomaterial? (Criticality level: ▲▲▲)

It is not sufficient to specify name or CAS-number for nanomaterials as they may cover a wide range of nanomaterials with chemical composition as the only common denominator. Nanomaterials should not only be identified from their chemical composition but also from other physico-chemical properties such as primary particle sizes, shape, crystal structure, and presence of coatings. Differences in these properties can prevent read across between two nanomaterials with the same chemical composition. Therefore, the information given in the study documentation must be consulted, where information should be available that clearly and unambiguously identifies the

nanomaterial for which the ecotoxicity data has been established. The detail of this information should make it possible to compare the specific tested substance with the substance for which e.g. the risk assessment is to be performed. If a specific tradename is given, if the material is a reference material (such as those provided by the National Institute of Standards and Technology (NIST, 2016)) or a representative industrial nanomaterial (such as those in the Joint Research Centre Nanomaterials Repository (JRC, 2016)) additional information can be obtained from the producer/supplier.

The following parameters are considered to be the minimum characterisation requirements that will allow for unambiguous substance identification:

- core chemical composition

- purity

- primary particle size (measured)

- shape

- crystal structure (if relevant)

- radiolabelled (if used)

- surface chemistry (e.g., coating material, functionalization if used)

In addition, information on surface reactivity and/or surface charge in the test media may be relevant as well as (photo)catalytic activity for some nanomaterials. Characterisation should be done using methods that are appropriate for nanomaterials. For aquatic ecotoxicity tests of some nanomaterials (like ZnO, Ag, and CuO), the water solubility of metal ions and related dissolution rates are parameters of high importance. To evaluate the reliability of the studies, information on dissolution of such nanomaterials must be provided. If the information is insufficient (generic and e.g. reporting only the chemical composition) the study is considered to be "not reliable" (nR3).

6 - Is the purity of the tested nanomaterial reported? This includes information on synthesis by-products as well as synthesis catalysts and presence of other crystalline forms of the substance. And/or, is the source of the test substance trustworthy? (Criticality level: ▲▲)

As for conventional chemicals it is important that tested nanomaterials are of high purity (for example listed as analytical grade, ACS reagent, extra pure grade, impurities typically in the range of 0.01-1%). Impurities may be formed as a result of synthesis. For example, carbon nanotubes (CNTs) may contain metal impurities that will contribute to the overall toxicity (Petersen & Henry 2012; Petersen et al. 2014). Also, different crystalline forms of the substance may be formed during the synthesis resulting in an unintended fraction of another crystalline form. This is for example the case for TiO_2 where the anatase form may contain fractions of rutile and amorphous particles. If, based on expert judgement, it is evaluated that the tested nanomaterial is likely to contain impurities, but if this is not reported and addressed in the study report, the reliability of the study should be lowered. It should be noted that regulatory testing (e.g. under REACH) requires the testing of the nanomaterial as put on the market, and hence not necessarily the analytical grade nanomaterial. In such cases it is still important not to erroneously attribute the effect to the nanomaterial in itself if effects are in fact caused by e.g. metal impurities. This could be addressed in the study for example by testing a leachate/filtrate of the nanomaterial (Petersen et al., 2014).

7 - If a formulation is used or if impurities or coatings are present: Do other ingredients in the formulation, the impurities or the coatings exert an effect? Is the amount of nanomaterial in the formulation known? (Criticality level: ▲/▲▲)

If a formulation of a nanomaterial is used, if impurities are reported or known to be present, or if the nanomaterial is coated (for example to stabilise it in a dispersion), the potential influence hereof on the ecotoxicity should be addressed. This should clarify if the observed effects can be partly or fully attributed to the formulation, impurities, and/or the coating material. If such information is not provided, then the reliability of the study should be lowered. For uncoated nanomaterials with a

high purity this is however not a critical criterion since effects are unlikely to arise from presence of impurities or coating material.

10 - Is the experimental system appropriate for the tested nanomaterial, taking into account its physico-chemical characteristics? Were measures taken to minimise losses and keep exposure stable (in quantitative and qualitative terms) throughout the duration of the test? (Criticality level: ▲▲)

As for conventional chemicals, the experimental system should be appropriate for the test substance in terms of e.g. vessel materials and use of methods that prevent loss of the test substance from the system (see Moermond et al. 2016). While glass is the preferred material for many conventional soluble chemicals this may not apply to nanomaterials. The test vessel should ideally be selected to minimise sorption or adherence of the nanomaterial to the surface of the container.

For nanomaterials it is often observed that the concentration in the water phase will decrease significantly during an aquatic ecotoxicity test due to particle aggregation/agglomeration and subsequent settling in the test vials. Stirring, shaking, use of flow through systems, and frequent media renewal are all ways of countering the loss of nanomaterials from the water phase. The overall setup of the experimental system as well, as any measures taken to avoid loss of the nanomaterial should of course, should of course be compatible with test organism requirements (e.g. light/dark conditions, exposure route etc.). If the experimental system is considered not appropriate, the study should be considered "not reliable" (nR3).

12 –If a dispersant/stabiliser/solvent is used, is the dispersant/stabilizer/solvent within the appropriate range and is a dispersant/stabiliser/solvent control included? (Criticality level: ▲/▲▲)

When preparing a test suspension the aim is not to dissolve the nanomaterial but rather to achieve a stable and homogeneous dispersion. To this end it can be necessary to add a dispersant/stabiliser/solvent of synthetic or natural origin. In addition, some nanomaterials are synthesized with a stabilizer, for example in the form of a capping agent, and thus by default present

in the test system. If a dispersant/stabiliser/solvent has been used the potential positive or adverse effects of this substance on the overall ecotoxicity should be addressed. If such information is not provided the reliability of the study should be lowered.

13 - Is a correct spacing between exposure concentrations applied? (Criticality level: ▲)

Defining the actual exposure concentrations is not trivial for ecotoxicity testing of nanomaterials. Firstly, there may be a correlation between nanomaterial concentration and agglomeration behaviour (Baalousha et al., 2016), meaning that exposure changes qualitatively along the concentration-axis in a concentration-response curve. Secondly, it has been suggested that concentration may not (alone) be sufficient as a dose-descriptor and other dose-metrics (such as surface area) have been proposed (van Hoecke et al., 2009). It may therefore be very challenging to even define a 'correct spacing' between exposure concentrations. This criterion is therefore considered not to be critical for the reliability of ecotoxicity studies of nanomaterials. Expert judgement should however be applied to ensure that the exposure concentrations are chosen in a reasonable manner.

15 - Have analyses been performed to verify exposure, e.g. concentrations and physicochemical transformations of the nanomaterial over the duration of the test? In case of unstable exposure, has this been accounted for in the data interpretation? Is the nanomaterial dispersion procedure described in detail? (Criticality level: ▲ ▲ ▲)

This criterion is linked to criteria 10 and 13 above. Achieving a stable nanomaterial test suspension is a challenging task as many nanomaterials tend to agglomerate and settle in aqueous solutions and especially in test media.

It is known that even small changes to the sample preparation and test methodology can have a great impact on the nanomaterial behaviour in the test system and resulting effects (Cupi et al.,

2016). Important factors include media composition, ionic strength, and pH. Also, it is important to know exactly how the test dispersion was prepared (including information on stock dispersion: nanomaterial nominal concentration, sonication time, sonication energy, volume and material of suspension vessel, pre-wetting steps, addition of dispersants etc.) as this will directly influence the characteristics of the test dispersion and hence the subsequent exposure conditions (Hartmann et al. 2015).

The introduction of a test organism is known to further destabilise the test suspension, leading to unstable exposure conditions during the tests (Hartmann et al., 2013). As mentioned it is therefore important to monitor nanomaterial exposure over the duration of the test, quantitatively as well as qualitatively. Important parameters to monitor over the duration of the test include:

- agglomeration state

- particle size distribution

- surface charge/zeta potential

- ion release

Preferably this information should be available for all tested concentrations and measured at regular intervals. For a short-term test this should be done as a minimum at the beginning and end of the test. Appropriate intervals should be established based on particle characteristics and behaviour in the test system. If for example agglomeration is known (or expected) to occur quickly then shorter intervals are required. For very stable suspensions longer intervals can be used. Also, in accordance with OECD guidance (2012) the characterization should be carried out for the nanomaterial as dry powder, in test media, in stock suspension and in the presence of test organisms using two or more analytical methods. All measurements and characterisation should be done using methods that are appropriate for the tested nanomaterials and results reported in relevant dose metrics.

For conventional chemicals the OECD ecotoxicity test guidelines operates with an acceptable limit of concentration variation of max ±20% during a test (Petersen et al. 2015). If the test concentration varies more than ±20% over the duration of the test this excludes the use of nominal concentrations and initial concentrations as a measure of exposure. Instead concentrations should be based on measurements throughout the test. Discussions are ongoing if such criteria are relevant for tests with nanomaterials (Petersen et al., 2015). So far there is no clear consensus on this matter. Nonetheless it is important that exposure concentrations are monitored (quantitatively and qualitatively) throughout the duration of the test and that its impact on observed effects is discussed.

16 - Is the biomass loading of the organisms in the test system within the appropriate range (< 1 g/L)? (Criticality level: ▲▲)

For nanomaterials physical 'overload' phenomena are sometimes observed and it is thus important to evaluate if the effects on the organisms is caused by such testing artefacts (Sørensen et al., 2015). Overload can result physical effects, for example with particles adhering to the surface of the organisms, causing impaired movement, or blockages of swimming appendages leading to increased energy cost for maintenance (Khan et al., 2017) . This requires an expert evaluation of the nanomaterial loading possibly combined with images of the biomass in the test system. A clear indication of testing artefacts caused by a high nanomaterial loading will negatively impact the reliability of the study.

19 - Is a concentration response curve observed? Is the response statistically significant? (▲/▲▲)

Regarding concentration response curve: see explanation to criterion 13. There may be inherent characteristics on the nanomaterial test system which impedes the establishment of concentration response curves. However, whether the observed responses are statistically significant is considered to be important for study reliability. Hence, the criticality of this criterion depends on the specific

type of test setup and output data. Expert judgement should be applied to evaluate the test results and the statistical significance of the observed response.

3.2. Relevance evaluation

Descriptions of the nanoCRED relevance categories are presented in Table 3, and the full list of relevance criteria is presented in Table 4. Nano-specific guidance to criterion 2 and 4 are provided since these criteria relate to relevant exposure route, endpoints, and mode of action for the tested substance (i.e. nanomaterial). As opposed to reliability evaluation we do not propose a weighing of the relevance criteria as this will be context specific. However, if desired, the evaluator can, based on expert judgment, weigh each relevance criterion as either "critical", "important" or "of minor importance". If doing so, clear justifications should be provided to enable transparent understanding of the evaluation.

Table 3. *Relevance categories (adapted from Moermond et al. 2016).*

Score	Description
nC1	**Relevant without restrictions:** The study is relevant for the purpose for which it is evaluated.
nC2	**Relevant with restrictions:** The study has limited relevance for the purpose for which it is evaluated.
nC3	**Not relevant**: The study is not relevant for the purpose for which it is evaluated.
nC4	**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 parameter.

Table 4. Relevance criteria (adopted from Moermond et al., 2016). Nano-specific guidance is available in the main text for criteria 2 and 4. For general explanatory guidance text on how to interpret these criteria see SI.

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 sub-stance?
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?

3.2.1. NanoCRED relevance criteria and guidance material

2 - Are the organisms tested relevant for the tested substance?

For studies on nanomaterials consideration may be given to the susceptibility of the test organism to particle exposure with regard to particle uptake and biological effects. For example, the normal feeding strategy of the organism may be considered (such as food size range, uptake mechanisms and location in water column).

4 - Are the reported endpoints appropriate for the investigated effects or the mode of action of the test substance?

For nanomaterials with a known mode of effect, studies that investigate these effects may be considered more relevant. At present, modes of effect are still not well understood for nanomaterials although mechanisms linked to the production of reactive oxygen species has been proposed (Skjolding et al. 2016). Effects may also include uptake-related non-standard endpoints and be linked to the potential for translocation of the nanomaterial away from the point of initial uptake (such as the gut) into other parts of the organism. As a better scientific understanding of effects mechanisms develop, the way in which this criterion is evaluated may change.

4. **DISCUSSION**

Nanomaterials are in many ways different from conventional chemicals, for which the majority of current ecotoxicological test guidelines have been developed. Nanomaterials will be present in the ecotoxicity test systems as physical entities, they undergo physical transformation processes during the test, and they are expected to have different modes of action compared to for example narcotic and polar narcotic chemicals. The test system dynamics can create a 'black box' testing situation due to uncontrollable changes in the exposure of the test organisms over time (e.g., Sørensen & Baun, 2015, Cupi et al., 2016). This makes it challenging to appropriately interpret the test results. The field of nanoecotoxicology testing has advanced over the last decade and existing test methods, including test guideline methods, have been adapted and refined to gain more control over the test systems. As nanomaterials are likely to have different uptake and effect mechanisms there might also be a need to use non-guideline test methods. This creates a conflict with the current evaluation of ecotoxicity data for regulatory purposes, which largely favours standard studies (Ågerstrand et al., 2011) and provides no nano-specific guidance.

The evaluation method as presented in this paper is based on an approach developed by Moermond et al. (2016) for conventional chemicals and involves an adaptation of the Klimisch method (Klimisch et. al., 1997). By applying this evaluation method the reliability of guideline test methods and non-guideline test methods are evaluated on equal terms. Although the presented evaluation method gives clear guidance on the interpretation of the individual criteria it does not provide a 'plug-and-play' tool. Rather it provides a structured, transparent approach to assist expert judgement in the assessment of the regulatory adequacy of ecotoxicological studies of nanomaterials. This is a deliberate choice to ensure a flexible and case-specific implementation of the criteria for different regulatory purposes as well as ensuring the inclusion of expert judgement in the process. As the nanoCRED method only provide guidance on criteria where nano-specific considerations are warranted, it should always be combined with the CRED evaluation method (Moermond et al., 2016), where more general guidance is given. To facilitate this, a document where these two methods have been combined is provided in the Supporting Information to this paper. Using the presented criteria the evaluator has to set up a specific procedure for weighting and summarising the overall assessment of relevance and reliability. This can be done in a quantitative or qualitative manner. While the method was developed for aquatic (pelagic) ecotoxicity data it may be adapted to encompass soil and sediment tests. This would, however, require that additional exposure specific aspects and analytical challenges are considered in the reliability criteria and accompanying guidance.

A quantitative assessment can be achieved by assigning numerical values to express the level of criteria fulfilment on a scale going from 'the study did not address this criterion at all' to 'satisfactory fulfilment'. This numerical value could be expressed as '% criteria fulfilment' or via a pre-defined scoring system (e.g. 1-10) i.e. a continuous range or as discrete values. The criticality weighing can further be assigned a certain numerical value for 'not critical criteria' (▲), 'important criteria' (▲▲) and 'critical criteria' (▲▲▲), respectively. By multiplying the 'fulfilment score' with the

'criticality score' the total weighed scores can be summed up. Pre-defined criteria for this overall score can be formulated for assigning reliability and relevance scores (nR1-nR4 and C1-C4) to the studies.

Alternatively, the criteria can be used in a more qualitative manner as a general guidance for data evaluation without assigning quantitative scores to the study. The criteria may also be use by researchers as a checklist for ensuring transparent and complete reporting of study design, substance characterisation and results of peer-reviewed ecotoxicity studies for nanomaterials.

After evaluating both the reliability and relevance of a study the study's overall adequacy for regulatory purposes can be assessed. This can be done according to the approach summarised in Figure 1.
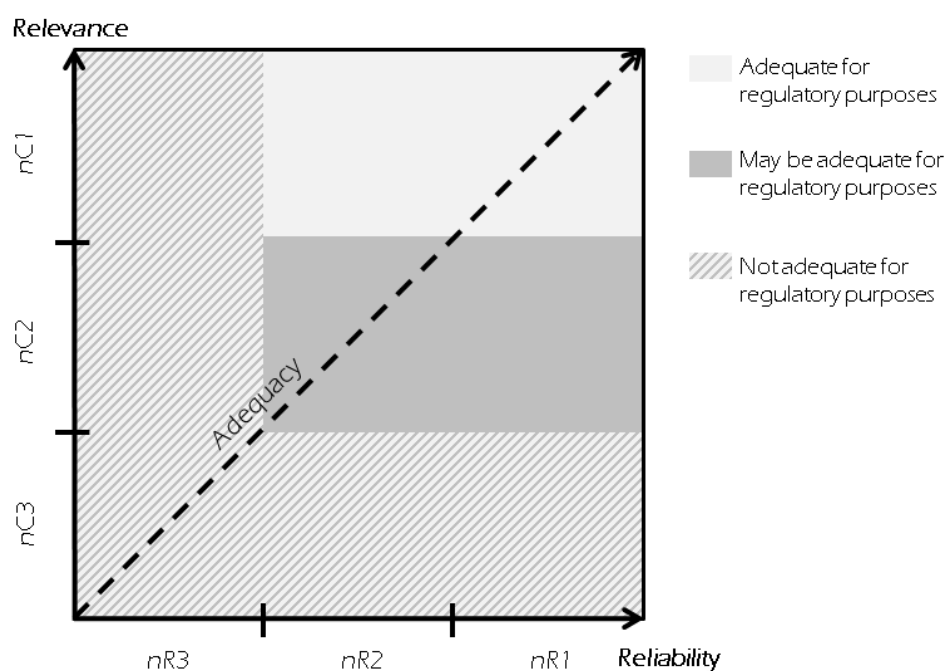

Figure 1. Proposed two-dimensional approach to summarize the adequacy of ecotoxicity data in risk assessment of nanomaterials based on data reliability and data relevance. nR1-nR3 refers to the reliability categorisation of the study and nC1-nC3 to the relevance categorisation. Adapted from Moermond et al. (2016).

In this approach a study is considered 'adequate for regulatory purposes' if it is reliable without restrictions (nR1) OR reliable with restrictions (nR2) AND relevant without restrictions (nC1). This type of studies would for example be used as key studies to derive PNEC values for risk assessment purposes. A study 'may be adequate for regulatory purposes' if it is reliable without restrictions (nR1) OR reliable with restrictions (nR2) AND relevant with restrictions (nC2). This type of studies could for example be considered as supporting studies in a risk assessment – or could be considered for PNEC derivations in case limited data are available. Any study that is considered 'not relevant' (nR3) AND/OR 'not reliable' (nC3) OR 'not assignable' (nR4/nC4) is considered 'not adequate for regulatory purposes'

One should be aware that there is a difference between scientifically valid studies and studies valid for regulatory purposes. A trade-off between reliability and relevance is often present in studies found adequate for regulatory decision-making due to the need to standardize and control experimental conditions to increase the reproducibility of test results (Wickson et al. 2014). This increased reliability often undermines the environmental realism of the tests and this does to a certain extend affect the relevance of the test results for dose-response assessments. While this realism-control dilemma has been recognized for conventional chemicals in the REACH guidance and dealt with by using extrapolation methods to estimate PNECs, this issue has not received much attention for nanomaterials (Baun et al. 2009). When nanomaterials are tested in standard ecotoxicity tests the validity criteria of stable exposure conditions during incubation are challenged by the behaviour of nanomaterials the test systems. The links between transformed states of nanomaterials (e.g. dissolved or agglomerate forms) and biological effects are at present unknown, but transformations during testing most certainly influence the test results (Sørensen & Baun, 2015; Thit et al., 2016). The framework for evaluating the regulatory adequacy of nanomaterial ecotoxicity studies as presented in this article provides scientifically based criteria to assist regulatory decision-

making. As science develops, it can be expected that also the criteria will change, e.g. due to higher demand for characterization during testing or increased environmental realisms of test results. Also, the formation of so-called corona around nanomaterials during ecotoxicity has been explored in recent years. It is formed as the result of interaction between nanomaterials and media constituents/organisms (Nasser & Lynch, 2014), and may significantly influence the bioavailability and toxicity of particles. It is therefore at present hard to give firm recommendations on how this should be taken into account in the reliability criteria. In other words, the presented framework is not intended to be static, but will hopefully develop according to the state-of-knowledge for ecotoxicity of nanomaterials. The online tool SciRAP (Science in Risk Assessment and Policy) (www.scirap.org) can be used to access the nanoCRED evaluation method, including future updates. SciRAP is a web-based resource developed specifically to assist in the reporting and evaluation of toxicity and ecotoxicity, including nanoecotoxicity, studies (Molander et al. 2015.

5. CONCLUSION

In this paper a method for evaluation of the reliability and relevance of nanomaterial ecotoxicity data for regulatory purposes is presented. It includes a range of nano-specific criteria together with detail guidance that will help risk assessors when deciding on if and how a study should be used in a regulatory context. The approach is applicable to both guideline and non-guideline studies and thereby deviating significantly from the currently used Klimisch approach, which inherently favours the use of guideline tests performed according to GLP. This evaluation framework may contribute to improved risk assessments of nanomaterials by enabling a data selection focussed at regulatory reliability and relevance and at the same time promoting transparency in the evaluation process.

Conflict of interest

The authors declare no conflicts of interests.

Acknowledgments

Supporting information

Manual for practical use of NanoCRED

REFERENCES

Ågerstrand, Marlene, Magnus Breitholtz, and Christina Rudén. 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. *Environmental Sciences Europe,* 23 (1): 17. doi:10.1186/2190-4715-23-17.

Ågerstrand, Marlene, Linnéa Edvardsson, and Christina Rudén. 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. *Human and Ecological Risk Assessment: An International Journal,* 20(6), pp. 1427-1445. doi:10.1080/10807039.2013.854139.

Aitken, R.A., Bassan, A., Friedrichs, S., Hankin, S.M., Hansen, S.F., Holmqvist, J., Peters, S.A.K., Poland, C.A. and Tran, C.L., 2011. Specific Advice on Exposure Assessment and Hazard/Risk Characterisation for Nanomaterials under REACH (RIP-oN 3)-Final Project Report. European Commission. Available from: http://ec.europa.eu/environment/chemicals/nanotech/pdf/report_ripon3.pdf

Baalousha M, Sikder M, Prasad A, Lead J R, Merrifield R, Chandler G T. The concentration-dependent behavior of nanoparticles. *Environmental Chemistry* 13 (2016) 1-3

Baun, A., Hartmann, N.B., Grieger, K.D. and Hansen, S.F., 2009. Setting the limits for engineered nanoparticles in European surface waters–are current approaches appropriate? *Journal of Environmental Monitoring*, 11(10), pp.1774-1781.

Brinch, A, Hansen, SF, Hartmann, NB & Baun, A 2016, EU Regulation of Nanobiocides: Challenges in Implementing the Biocidal Product Regulation (BPR). *Nanomaterials*, vol 6, no. 2, 33., 10.3390/nano6020033

Card, J.W. and Magnuson, B.A., 2010. A method to assess the quality of studies that examine the toxicity of engineered nanomaterials. *International Journal of Toxicology*, 29(4), pp.402-410.

Cupi, D, Hartmann, NB & Baun, A 2016, Influence of pH and media composition on suspension stability of silver, zinc oxide, and titanium dioxide nanoparticles and immobilization of Daphnia magna under guideline testing conditions. *Ecotoxicology and Environmental Safety*, vol 127, pp. 144-152., 10.1016/j.ecoenv.2015.12.028

Cupi, D., Hartmann,N.B., Baun, A. (2015). The influence of natural organic matter and aging on suspension stability in guideline toxicity testing of ZnO, TiO2, and Ag nanoparticles with Daphnia magna. *Environmental Toxicology and Chemistry* 34(3):497-506.

DaNa. 2016. Literature Criteria Checklist. Methodology of selection, acquisition and evaluation of toxicological publications in the Project DaNa. Version 2016. Available from: http://nanopartikel.info/en/nanoinfo/methods (accessed 16th Deceber 2016)

Durda, Judi L., and Damian V. Preziosi. 2000. Data Quality Evaluation of Toxicological Studies Used to Derive Ecotoxicological Benchmarks. *Human and Ecological Risk Assessment* 6 (5): 747–65. doi:10.1080/10807030091124176.

European Chemicals Agency (ECHA). 2008. Guidance on information requirements and chemical safety assessment Chapter R.10: Characterisation of dose [concentration]-response for environment. Available from:

http://echa.europa.eu/documents/10162/13632/information_requirements_r10_en.pdf

European Chemicals Agency (ECHA) 2014. Guidance on Information Requirements and Chemical Safety Assessment Chapter R.7c: Endpoint specific guidance, Version 2.0 November 2014

European Chemicals Agency (ECHA) 2015. Guidance on Information Requirements and Chemical Safety Assessment Chapter R.7a: Endpoint specific guidance, Version 4.1 October 2015

European Chemicals Agency (ECHA) 2016. Guidance on Information Requirements and Chemical Safety Assessment Chapter R.7b: Endpoint specific guidance, Version 3.0 February 2016

European Commission (EC) 2006. Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC.

European Commission (EC) 2008. Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006.

European Commission (EC) 2012. NANO SUPPORT Project - Scientific technical support on assessment of nanomaterials in REACH registration dossiers and adequacy of available information on analysis and assessment. Final Report on analysis and assessment (Task I, step 3&4&5) and

options for adapting REACH (Task II, step 1). Available from :

http://ec.europa.eu/environment/chemicals/nanotech/pdf/jrc_report.pdf

European Union (EU), 2011. Commission recommendation of 18 October 2011 on the definition of nanomaterial (2011/696/EU). Official Journal L 275, 38–40

Hansen, S.F., Hjorth, R., Skjolding, L.M, Bowman, D.M., Maynard, A., Baun, A. (2017). A critical analysis of the Environmental Dossiers in the OECD Sponsorship Programme for Manufactured Nanomaterials. *Environ. Sci.: Nano*, 2017, 4, 282 DOI:10.1039/C6EN00465B

Hartmann, NB, Engelbrekt, C, Zhang, J, Ulstrup, J, Kusk, KO & Baun, A 2013, The challenges of testing metal and metal oxide nanoparticles in algal bioassays: titanium dioxide and gold nanoparticles as case studies. *Nanotoxicology*, vol 7, no. 6, pp. 1082-1094., 10.3109/17435390.2012.710657

Hartmann, N.B., Jensen, K.A., Baun, A., Rasmussen, K., Rauscher, H., Tantra, R., Cupi, D., Gilliland, D., Pianella, F. and Riego Sintes, J.M., 2015. Techniques and protocols for dispersing nanoparticle powders in aqueous media—Is there a rationale for harmonization? *Journal of Toxicology and Environmental Health, Part B*, 18(6), pp.299-326.

Hobbs, Dustin A, Michael St J Warne, and Scott J Markich. 2005. Evaluation of Criteria Used to Assess the Quality of Aquatic Toxicity Data. *Integrated Environmental Assessment and Management* 1 (3): 174–80.

Hund-Rinke, K., Baun, A., Cupi, D., Fernandes, T.F., Handy, R., Kinross, J.H., Navas, J.M., Peijnenburg, W., Schlich, K., Shaw, B.J. and Scott-Fordsmand, J.J., 2016. Regulatory ecotoxicity testing of

nanomaterials–proposed modifications of OECD test guidelines based on laboratory experience with silver and titanium dioxide nanoparticles. *Nanotoxicology*, 10(10), pp. 1442-1447

Joint Research Centre (JRC), 2017.ToxRTool - Toxicological data Reliability Assessment Tool. URL: https://eurl-ecvam.jrc.ec.europa.eu/about-ecvam/archive-publications/toxrtool (accessed 21nd February 2017)

Joint Research Center (JRC), 2016. JRC Nanomaterials Repository. URL: https://ec.europa.eu/jrc/en/scientific-tool/jrc-nanomaterials-repository (accessed 21nd February 2017)

Kase, R., Korkaric, M., Werner, I. and Ågerstrand, 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. *Environmental Sciences Europe*, 28(1), p.1.

Khan, F.R., Syberg, K., and Palmqvist A., 2017. Are Standardized Test Guidelines Adequate for Assessing Waterborne Particulate Contaminants? *Environmental Science & Technology* 51 (4), 1948-1950

Klimisch, H.-J, Andreae,M. and Tillmann, U. 1997. A Systematic Approach for Evaluating the Quality of Experimental Toxicological and Ecotoxicological Data. *Regulatory Toxicology and Pharmacology* 25 (1): 1–5. doi:10.1006/rtph.1996.1076.

Kühnel, D. and Nickel, C., 2014. The OECD expert meeting on ecotoxicology and environmental fate—towards the development of improved OECD guidelines for the testing of nanomaterials. *Science of the Total Environment*, 472, pp.347-353.

Küster, A., J. Bachmann, U. Brandt, I. Ebert, S. Hickmann, J. Klein-Goedicke, G. Maack, S. Schmitz, E. Thumm, and B. Rechenberg. 2009. Regulatory Demands on Data Quality for the Environmental Risk Assessment of Pharmaceuticals. *Regulatory Toxicology and Pharmacology* 55 (3): 276–80. doi:10.1016/j.yrtph.2009.07.005.

Lubinski, L., Urbaszek, P., Gajewicz, A., Cronin, M.T.D., Enoch, S.J., Madden, J.C., Leszczynska, D., Leszczynski, J. and Puzyn, T., 2013. Evaluation criteria for the quality of published experimental data on nanomaterials and their usefulness for QSAR modelling. *SAR and QSAR in Environmental Research*, 24(12), pp.995-1008.

Moermond, C., Kase, R., Korkaric, M. & Ågerstrand, M. 2016. CRED: Criteria for reporting and evaluating ecotoxicity data. *Environmental Toxicology and Chemistry*, 35(5), pp.1297–1309.

Molander, L., Ågerstrand, M., Beronius, A., Hanberg, A. and Rudén, C., 2015. Science in Risk Assessment and Policy (SciRAP): an online resource for evaluating and reporting in vivo (eco) toxicity studies. *Human and Ecological Risk Assessment: An International Journal*, 21(3), pp.753-762.

Nasser, F. and Lynch, I., 2016. Secreted protein eco-corona mediates uptake and impacts of polystyrene nanoparticles on Daphnia magna. *Journal of Proteomics*, 137, pp.45-51.

NIST (2016). Standard Reference Materials. URL: https://www-s.nist.gov/srmors/ (accessed 16th December 2016)

Notter, D. a., Mitrano, D. M., & Nowack, B. (2014). Are nanosized or dissolved metals more toxic in the environment? A meta-analysis. *Environmental Toxicology and Chemistry*, 33(12), 2733–2739. doi:10.1002/etc.2732

OECD. 2012. Guidance on Sample Preparation and Dosimetry for the Safety Testing of Manufactured Nanomaterials. OECD Environment, Health and Safety Publications. Series on the Safety of Manufactured Nanomaterials. ENV/JM/MONO(2012)/40.

Petersen, E.J., Diamond, S.A., Kennedy, A.J., Goss, G.G., Ho, K., Lead, J., Hanna, S.K., Hartmann, N.B., Hund-Rinke, K., Mader, B., Manier, N., Pandard, P., Salinas, E. & Sayre, P., 2015. Adapting OECD aquatic toxicity tests for use with manufactured nanomaterials: key issues and consensus recommendations. *Environmental Science & Technology*, vol 49, no. 16, pp. 9532-9547., 10.1021/acs.est.5b00997

Petersen, E.J. and Henry, T.B., 2012. Methodological considerations for testing the ecotoxicity of carbon nanotubes and fullerenes: review. *Environmental Toxicology and Chemistry*, 31(1), pp.60-72.

Petersen, E.J., Henry, T.B., Zhao, J., MacCuspie, R.I., Kirschling, T.L., Dobrovolskaia, M.A., Hackley, V., Xing, B. and White, J.C., 2014. Identification and avoidance of potential artifacts and misinterpretations in nanomaterial ecotoxicity measurements. *Environmental Science & Technology*, 48(8), pp.4226-4246.

Pettitt, M.E. and Lead, J.R., 2013. Minimum physicochemical characterisation requirements for nanomaterial regulation. *Environment International*, 52, pp.41-50.

Robinson, R.L.M., Lynch, I., Peijnenburg, W., Rumble, J., Klaessig, F., Marquardt, C., Rauscher, H., Puzyn, T., Purian, R., Åberg, C. and Karcher, S., 2016. How should the completeness and quality of curated nanomaterial data be evaluated? *Nanoscale*, *8*(19), pp.9919-9943.

Schneider, Klaus, Markus Schwarz, Iris Burkholder, Annette Kopp-Schneider, Lutz Edler, Agnieszka Kinsner-Ovaskainen, Thomas Hartung, and Sebastian Hoffmann. 2009. 'ToxRTool', a New Tool to Assess the Reliability of Toxicological Data. *Toxicology Letters* 189 (2): 138–44. doi:10.1016/j.toxlet.2009.05.013.

Selck, H., Handy, R.D., Fernandes, T.F., Klaine, S.J. and Petersen, E.J., 2016. Nanomaterials in the aquatic environment: A European Union–United States perspective on the status of ecotoxicity testing, research priorities, and challenges ahead. *Environmental Toxicology and Chemistry*, 35(5), pp.1055-1067.

Skjolding, L.M., Sørensen, S.N., Hartmann, N.B., Hjorth, R., Hansen, S.F. and Baun, A., 2016. A Critical Review of Aquatic Ecotoxicity Testing of Nanoparticles-The Quest for Disclosing Nanoparticle Effects. *Angewandte Chemie Int. Ed.,* 55, pp. 15224–15239.

Sørensen, S.N., Hjorth, R., Delgado, C.G., Hartmann, N.B. and Baun, A., 2015. Nanoparticle ecotoxicity—physical and/or chemical effects? *Integrated environmental assessment and management*, 11(4), pp.722-724.

Sørensen, S.N. & Baun, A. (2015). Controlling silver nanoparticle exposure in algal toxicity testing – a matter of timing. *Nanotoxicology*, 9, 201-209.

UN (United Nations). 2011. Globally Harmonized System of Classification and Labelling of Chemicals (GHS). Fourth revised edition. United Nations, New York and Geneva.

US Envirpnmental Protection Agency (USEPA) (2015). Regulatory Actions under TSCA Section 5. url: https://www.epa.gov/reviewing-new-chemicals-under-toxic-substances-control-act-tsca/regulatory-actions-under-tsca#5a3c (last aaccessed 16th December 2016)

Wickson, F., Hartmann, N.B., Hjorth, R., Hansen, S.F., Wynne, B. and Baun, A., 2014. Balancing scientific tensions. *Nature Nanotechnology*, 9(11), pp.870-870.

van Hoecke, K., Quik, J.T., Mankiewicz-Boczek, J., Schamphelaere, K.A.D., Elsaesser, A., Meeren, P.V.D., Barnes, C., McKerr, G., Howard, C.V., Meent, D.V.D. and Rydzynski, K., 2009. Fate and effects of CeO2 nanoparticles in aquatic ecotoxicity tests. *Environmental Science & Technology*, 43(12), pp.4537-4546.

Supporting information

Manual for the practical use of nanoCRED

NanoCRED: A transparent method to assess the regulatory adequacy of ecotoxicity data for nanomaterials – relevance and reliability revisited

NB Hartmann†*, M Ågerstrand‡, HCH Lützhøft†, A Baun†

†Department of Environmental Engineering, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark (nibh@env.dtu.dk)

‡ Department of Environmental Science and Analytical Chemistry, Stockholm University, SE-106 91 Stockholm, Sweden

*To whom correspondence may be addressed. E-mail: nibh@env.dtu.dk Telephone: +45-45251401; Fax: +31-30-2744401

Table of Contents

Introduction

The nanoCRED evaluation method was developed to guide risk assessors when evaluating nanoecotoxicity studies for regulatory purposes. It provides criteria and additional guidance material for evaluation of reliability as well as relevance. The nanoCRED evaluation method offers a structured approach that promotes transparency and thorough assessment.

Overview of the process of evaluating study adequacy

Figure S1 gives an overview of the process of evaluating the regulatory adequacy of nanoecotoxicity data. The individual steps are described in more details below. As shown, the evaluation comprises an evaluation of study reliability and an evaluation of study relevance. These two evaluations are subsequently combined in an overall evaluation of study adequacy.

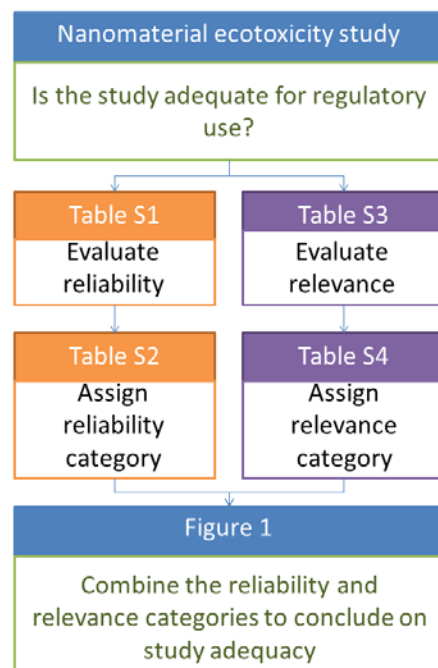**Figure S1.** *Illustration of the process of evaluating the regulatory adequacy of nanoecotoxicity data.* The approach include the following subsequent steps 1) evaluation of reliability, 2) assignment of reliability category (nR), 3) evaluation of relevance, 4) assignment of relevance category (nC) and 5) combining reliability and relevance evaluations into an overall evaluation of study adequacy.

Step 1: Evaluate reliability

Reliability is evaluated using 20 reliability criteria (Table S1). The reliability criteria are based on the CRED evaluation method (Moermond et al. 2016) with nano-specific considerations for criterion number 1-7, 10, 12-13, 15-16, and 19. The degree of fulfilment of each criterion is marked, either using "fulfilled", "partly fulfilled", or "not fulfilled", or by using a "fulfilment score", i.e. assigning numerical values (discrete values or a continuous range) on a scale going from 'the study did not address this criterion at all' to 'satisfactory fulfilment'.

To guide the risk assessor further, the 20 reliability criteria have been weighted using the following descriptors: "not critical (▲), 'important' (▲▲) and 'critical" (▲▲▲). This weighing represents the importance of each criterion when evaluating nanoecotoxicity studies.

As an option, there is also a possibility to quantitatively summarize the evaluation. This can for example be done by multiplying a 'fulfilment score' with the 'weighing' score. If doing so, it is important to remember that a number can be a supplement to a qualitative summarising but it cannot replace the decision support that a detailed description of a study's strengths and weaknesses provides.

Table S1. *Reliability criteria for nanoecotocicity data. The criteria are adapted and modified from the CRED evaluation method (Moermond et al. 2016) taking into account nano-specific concerns. Additional explanatory guidance on how to interpret these criteria is available in the main text below. The criteria are weighted using the following descriptors: critical for study reliability (▲▲▲), important for study reliability (▲▲), of minor importance for study reliability, but may support study evaluation (▲), the importance of these criteria depend on the specific test and/or nanomaterial properties (▲/▲▲).*

Number	Weighing	Criterion
		General information
		Before evaluating the test, check the physico-chemical characteristics of your nanomaterial (handbooks/general sources/scientific literature).
		Test setup
1	▲	Is a guideline method (e.g., OECD/ISO) or modified guideline used?
2	▲	Is the test performed under GLP conditions?
3	▲/▲▲	If applicable, are validity criteria fulfilled (e.g. control survival, growth)?
4	▲▲▲	Are appropriate controls performed (e.g. dispersant control, metal ion control, larger than nano-sized (bulk) particles of the same chemical composition, negative and positive control)?

Test Compound

5	▲ ▲ ▲	Is the tested nanomaterial appropriately identified with name or chemical identifier (e.g., CAS-number)? Are nanomaterial characteristics reported that allow for a clear identification of the tested nanomaterial (e.g. particle size, shape, particle size distribution, surface area, coating)? Are test results reported for the relevant nanomaterial?
6	▲ ▲	Is the purity of the tested nanomaterial reported? This includes information on synthesis by-products as well as synthesis catalysts and presence of other crystalline forms of the substance. And/or, is the source of the test substance trustworthy?
7	▲/▲ ▲	If a formulation is used or if impurities or coatings are present: Do other ingredients in the formulation, the impurities or the coatings exert an effect? Is the amount of nanomaterial in the formulation known?

Test Organism

8	▲ ▲	Are the organisms well described (e.g. scientific name, sex, 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 the nanomaterial or other unintended stressors?

		Exposure Conditions
10	▲▲	Is the experimental system appropriate for the tested nanomaterial, taking into account its physico-chemical characteristics? Were measures taken to minimise losses and keep exposure stable (in quantitative and qualitative terms) throughout the duration of the test?
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	▲/▲▲	If a dispersant/stabiliser/solvent is used, is the dispersant/stabilizer/solvent within the appropriate range and is a dispersant/stabiliser/solvent control included?
13	▲	Is a correct spacing between exposure concentrations applied?
14	▲▲▲	Is the exposure duration defined?
15	▲▲▲	Have analyses been performed to verify exposure, e.g. concentrations and physico-chemical transformations of the nanomaterial over the duration of the test? In case of unstable exposure, has this been accounted for in the data interpretation? Is the nanomaterial dispersion procedure described in detail?
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)?

Explanation of the reliability criteria (numbers refer to Table S1)

*The guidance below is a combination of the general **CRED** guidance (from Moermond et al. 2016) and the guidance specific for nanomaterial studies from the **nanoCRED** method (Hartmann et al., 2017).*

1 - Is a standard method (e.g., OECD, ISO) or modified standard used? (Criticality level: ▲)

CRED: "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" (from Moermond et al., 2016).

nanoCRED: As for conventional chemicals the use of standard methods for ecotoxicity testing of nanomaterials is not in itself a guarantee for study reliability. Use of standard methods may ensure better and more complete documentation but does not per se ensure e.g. an appropriate study design or correct interpretation of test results (Moermond et al. 2016). This is especially true for nanomaterials. This is due to the fact that current guidelines were developed mainly for soluble chemicals and do not take into account the particulate nature of nanomaterials. Until standard methods are available explicitly for testing of nanomaterials, the use of standard methods may therefore result in lower reliability compared to non-standard tests – or modified guideline tests. A well-considered modification of a standard method is preferred over an uncritical use of a standard method that does not account for the nano-specific physical and chemical properties. ***The use of standard guidelines is therefore considered to be of minor importance for reliability per se.***

2 - Is the test performed under GLP conditions? (Criticality level: ▲)

CRED: "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" from Moermond et al., 2016).

nanoCRED: As for standard guidelines, GLP helps to ensure reproducibility and transparency but is not in itself a guarantee for reliability. The use of GLP is therefore considered to be a criterion of minor importance for study reliability.

3 - If applicable, are validity criteria fulfilled (e.g. control survival, growth)? (Criticality level: ▲/▲ ▲)

CRED: "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" (from Moermond et al., 2016).

nanoCRED: This criterion especially applies to studies that follow guidelines or modified guidelines containing requirements regarding validity criteria. However, any study with control group mortality may indicate a test set-up problem. For

nanomaterials modifications of the guidelines are often needed to accommodate the different behaviour of the nanoparticles compared to conventional soluble chemicals, for which most of the currently available test guidelines were originally intended. For modified guidelines the original validity criteria may not be relevant or possible to achieve. In these cases, expert judgement is required. For a description of general issues related to this criterion see Moermond et al. (2016).

4 - Are appropriate controls performed (e.g., dispersant control, metal ion control, larger than nano-sized (bulk) particles of the same chemical composition, negative and positive control)? (Criticality level: ▲▲▲)

CRED: "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" (from Moermond et al., 2016).

nanoCRED: For conventional chemicals it is common to use solvents to dissolve poorly soluble compounds. For nanomaterials it is not feasible and/or relevant to dissolve the particles. Instead the aim is rather to achieve a stable and homogeneous suspension (Hartmann et al., 2015). For example, when testing silver nanoparticles, the aim is not to dissolve the particles into ions and ion complexes but rather to test silver in its nano-particulate form. Stable and homogeneous dispersions can sometimes be achieved by adding a dispersant or a stabiliser to the media. This could be a substance that adsorbs to the surface of the particle

and causes electrostatic or steric stabilisation. It is important to include dispersant controls to ensure that this substance does not in itself cause toxic effects.

If a dispersant has been used, but no information is provided regarding its ecotoxicity, the study is considered to be "not reliable" (nR3)*.*

Some nanomaterials will dissolve partly in aqueous testing media during testing. This is for example the case for some silver and zinc oxide nanoparticles in standard algal and daphnia media (e.g. Cupi et al. 2016; Sørensen & Baun 2015). Other nanomaterials, like TiO2, will not undergo dissolution under the testing conditions of ecotoxicity test. For nanomaterials which release metal ions it is important that the study also addresses the metal ion contribution to the observed toxicity. This is often done by including metal ion controls and monitoring the ion release from the nanomaterial. The key to data interpretation is then to evaluate the particle as well as the ion contributions to the observed ecotoxicological effects. Expert judgement is therefore needed to evaluate if the study is lacking critical information regarding metal ion controls.

Nanomaterials are known to introduce certain test artefacts leading to potential misinterpretation of results (Selck et al., 2016; Sørensen et al., 2015). For this reason other types of controls, may be relevant to address e.g. effects caused by endotoxins (as a material impurity), indirect effects (such as effects on nutrient or light availability), photoactivity, fluorescence or absorbance properties (influencing assay measurement) (Selck et al., 2016; Petersen et al., 2014). These and other potential control experiments are described in detail in Petersen et al. (2014).

A negative control refers to test samples in pure media without the presence of nanomaterials, dispersants etc. Mortality in the negative controls may indicate some problems with the study reliability and result in a lower reliability score (nR2 or nR3) subject to expert judgement. Positive controls are useful to check the sensitivity of the test organisms and test performance and are often included as validity criteria in standard guideline tests (see criterion 3). If positive controls show an abnormal response this should be taken into account when evaluating the reliability of a study.

5 - Is the tested nanomaterial identified with name or CAS-number? Are nano-material characteristics reported that allow for a clear identification of the tested nanomaterial (e.g. particle size, shape, particle size distribution, surface area, coating)? Are test results reported for the relevant nanomaterial? (Criticality level: ▲ ▲ ▲)

CRED: *"*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 H2O) 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)" (from Moermond et al., 2016).

nanoCRED: It is not sufficient to specify name or CAS-number for nanomaterials as they may cover a wide range of nanomaterials with chemical composition as the only common denominator. Nanomaterials should not only be identified from their chemical composition but also from other physico-chemical properties such as primary particle sizes, shape, crystal structure, and presence of coatings. Differences in these properties can prevent read across between two nano-materials with the same chemical composition. Therefore, the information given in the study documentation must be consulted, where information should be available that clearly and unambiguously identifies the nanomaterial for which the ecotoxicity data has been established. The detail of this information should make it possible to compare the specific tested substance with the substance for which e.g. the risk assessment is to be performed. If a specific tradename is given, if the material is a reference material (such as those provided by the National Institute of Standards and Technology (NIST, 2016)) or a representative industrial nanomaterial (such as those in the Joint Research Centre Nanomaterials Reposi-tory (JRC, 2016)) additional information can be obtained from the produc-er/supplier.

The following parameters are considered to be the minimum characterisation requirements that will allow for unambiguous substance identification:

- core chemical composition,

- purity,

- primary particle size (measured),

- shape,

- crystal structure (if relevant),

- radiolabelled (if used),

- surface chemistry (e.g., coating material, functionalization if used),

In addition, information on surface reactivity and/or surface charge in the test media may be relevant as well as (photo)catalytic activity for some nanomaterials. Characterisation should be done using methods that are appropriate for nanomaterials. For aquatic ecotoxicity tests of some nanomaterials (like ZnO, Ag, and CuO), the water solubility of metal ions and related dissolution rates are parameters of high importance. To evaluate the reliability of the studies, information on dissolution of such nanomaterials must be provided.

If the information is insufficient (generic and e.g. reporting only the chemical composition) the study is considered to be "not reliable" (nR3).

6 - Is the purity of the tested nanomaterial reported? This includes information on synthesis by-products as well as synthesis catalysts and presence of other crystalline forms of the substance. And/or, is the source of the test substance trustworthy? (Criticality level: ▲ ▲)

CRED: "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" (from Moermond et al., 2016).

nanoCRED: As for conventional chemicals it is important that tested nanomaterials are of high purity (for example listed as analytical grade, ACS reagent, extra pure grade, impurities typically in the range of 0.01-1%). Impurities may be formed as a result of synthesis. For example, carbon nanotubes (CNTs) may contain metal impurities that will contribute to the overall toxicity (Petersen & Henry 2012; Petersen et al. 2014). Also, different crystalline forms of the substance may be formed during the synthesis resulting in an unintended fraction of another crystalline form. This is for example the case for TiO2 where the anatase form may contain fractions of rutile and amorphous particles. If, based on expert judgement, it is evaluated that the tested nanomaterial is likely to contain impurities, but if this is not reported and addressed in the study report, the reliability of the study should be lowered. It should be noted that regulatory testing (e.g. under REACH) requires the testing of the nanomaterial as put on the market, and hence not necessarily the analytical grade nanomaterial. In such cases it is still important not to erroneously attribute the effect to the nanomaterial in itself if effects are in fact caused by e.g. metal impurities. This could be addressed in the study for example by testing a leachate/filtrate of the nanomaterial (Petersen et al., 2014).

7 - If a formulation is used or if impurities or coatings are present: Do other ingredients in the formulation, the impurities or the coatings exert an effect? Is the amount of nanomaterial in the formulation known? (Criticality level: ▲/▲ ▲)

CRED: "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 as-

sessment, 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" (from Moermond et al., 2016).

nanoCRED: If a formulation of a nanomaterial is used, if impurities are reported or known to be present, or if the nanomaterial is coated (for example to stabilise it in a dispersion), the potential influence hereof on the ecotoxicity should be addressed. This should clarify if the observed effects can be partly or fully attributed to the formulation, impurities, and/or the coating material. If such information is not provided then the reliability of the study should be lowered. For uncoated nanomaterials with a high purity this is however not a critical criterion since effects are unlikely to arise from presence of impurities or coating material.

8 - Are the organisms well described (e.g., scientific name, sex, weight, length, growth, age/life stage, strain/clone, gender if appropriate)? (Criticality level: ▲▲)

CRED: "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" (from Moermond et al. 2016).

nanoCRED: no nano-specific guidance for this criterion.

9 - Are the test organisms from a trustworthy source and acclimatized to test conditions? Have the organisms not been pre-exposed to the nanomaterial or other unintended stressors? (Criticality level: ▲▲▲)

CRED: "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" (from Moermond et al. 2016).

nanoCRED: no nano-specific guidance for this criterion.

10 - Is the experimental system appropriate for the tested nanomaterial, taking into account its physico-chemical characteristics? Were measures taken to mini-mise losses and keep exposure stable (in quantitative and qualitative terms) throughout the duration of the test? (Criticality level: ▲▲)

CRED: 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" (from Moermond et al., 2016).

nanoCRED: As for conventional chemicals, the experimental system should be appropriate for the test substance in terms of e.g. vessel materials and use of methods that prevent loss of the test substance from the system (see Moermond et al. 2016). While glass is the preferred material for many conventional soluble chemicals this may not apply to nanomaterials. The test vessel should ideally be selected to minimise sorption or adherence of the nanomaterial to the surface of the container.

For nanomaterials it is often observed that the concentration in the water phase will decrease significantly during an aquatic ecotoxicity test due to particle aggregation/agglomeration and subsequent settling in the test vials. Stirring, shaking, use of flow through systems, and frequent media renewal are all ways of countering the loss of nanomaterials from the water phase. The overall setup of the experimental system, as well as any measures taken to avoid loss of the nanomaterial, should of course be compatible with test organism requirements (e.g. light/dark conditions, exposure route etc.).

If the experimental system is considered not appropriate, the study should be considered "not reliable" (nR3).

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? (Criticality level: ▲▲▲)

CRED: 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 (from Moermond et al. 2016).

nanoCRED: no nano-specific guidance for this criterion.

12 –If a dispersant/stabiliser/solvent is used, is the dispersant/stabilizer/solvent within the appropriate range and is a dispersant/stabiliser/solvent control included? (Criticality level: ▲/▲▲)

CRED: "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)" (from Moermond et al., 2016).

nanoCRED: This criterion is linked to criterion 4. When preparing a test suspension the aim is not to dissolve the nanomaterial but rather to achieve a stable

and homogeneous dispersion. To this end it can be necessary to add a dispersant/stabiliser/solvent of synthetic or natural origin. In addition, some nanomaterials are synthesized with a stabilizer, for example in the form of a capping agent, which is thus by default present in the test system. If a dispersant/stabiliser/solvent has been used the potential positive or adverse effects of this substance on the overall ecotoxicity should be addressed. Expert judgement should be applied to evaluate if concentrations are within an appropriate range. If the provided information is inadequate to make such an evaluation then the reliability of the study should be lowered.

13 - Is a correct spacing between exposure concentrations applied? (Criticality level: ▲)

CRED: "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 (=√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" (from Moermond et al., 2016).

nanoCRED: Defining the actual exposure concentrations is not trivial for ecotoxicity testing of nanomaterials. Firstly, there may be a correlation between nanomaterial concentration and agglomeration behaviour (Baalousha et al. 2016), meaning that exposure changes qualitatively along the concentration-axis in a concentration-response curve. Secondly, it has been suggested that concentration may not (alone) be sufficient as a dose-descriptor and other dose-metrics (such as surface area) have been proposed (van Hoecke et al., 2009). It may therefore be very challenging to even define a 'correct spacing' between exposure concentrations. This criterion is therefore considered not to be critical for the reliability of ecotoxicity studies of nanomaterials. Expert judgement should however be applied to ensure that the exposure concentrations are chosen in a reasonable manner.

14 -Is the exposure duration defined? (Criticality level: ▲ ▲ ▲)

CRED: 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 (from Moermond et al. 2016).

nanoCRED: no nano-specific guidance for this criterion.

15 - Have analyses been performed to verify exposure, e.g. concentrations and physicochemical transformations of the nanomaterial over the duration of the test? In case of unstable exposure, has this been accounted for in the data interpretation? Is the nanomaterial dispersion procedure described in detail? (Criticality level: ▲ ▲ ▲)

CRED: "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" (from Moermond et al., 2016).

nanoCRED: This criterion is linked to criteria 10 and 13 above. Achieving a stable nanomaterial test suspension is a challenging task as many nanomaterials tend to agglomerate and settle in aqueous solutions and especially in test media.

It is known that even small changes to the sample preparation and test methodology can have a great impact on the nanomaterial behaviour in the test system and resulting effects (Cupi et al., 2016). Important factors include media composition, ionic strength, and pH. Also, it is important to know exactly how the test dispersion was prepared (including information on stock dispersion: nanomaterial nominal concentration, sonication time, sonication energy, volume and material of suspension vessel, pre-wetting steps, addition of dispersants etc.) as this will directly influence the characteristics of the test dispersion and hence the subsequent exposure conditions (Hartmann et al. 2015).

The introduction of a test organism is known to further destabilise the test suspension, leading to unstable exposure conditions during the tests (Hartmann et al., 2013). As mentioned it is therefore important to monitor nanomaterial expo-

sure over the duration of the test, quantitatively as well as qualitatively. Important parameters to monitor over the duration of the test include:

- agglomeration state

- particle size distribution

- surface charge/zeta potential

- ion release

Preferably this information should be available for all tested concentrations and measured at regular intervals. For a short-term test this should be done as a minimum at the beginning and end of the test. Appropriate intervals should be established based on particle characteristics and behaviour in the test system. If for example agglomeration is known (or expected) to occur quickly then shorter intervals are required. For very stable suspensions longer intervals can be used. Also, in accordance with OECD guidance (2012) the characterization should be carried out for the nanomaterial as dry powder, in test media, in stock suspension and in the presence of test organisms using two or more analytical methods. All measurements and characterisation should be done using methods that are appropriate for the tested nanomaterials and results reported in relevant dose metrics.

For conventional chemicals the OECD ecotoxicity test guidelines operates with an acceptable limit of concentration variation of max ±20% during a test (Petersen et al. 2015). If the test concentration varies more than ±20% over the duration of the test this excludes the use of nominal concentrations and initial concentrations as a measure of exposure. Instead concentrations should be based on measurements throughout the test. Discussions are ongoing if such criteria are relevant for tests with nanomaterials (Petersen et al., 2015). So far there is no clear consensus on this matter. Nonetheless it is important that exposure concentrations are monitored (quantitatively and qualitatively) throughout the duration of the test and that its impact on observed effects is discussed.

16 - Is the biomass loading of the organisms in the test system within the appropriate range (< 1 g/L)? (Criticality level: ▲ ▲)

CRED: "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 KOW 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" (from Moermond et al., 2016).

nanoCRED: For nanomaterials physical 'overload' phenomena are sometimes observed and it is thus important to evaluate if the effects on the organisms is caused by such testing artefacts (Sørensen et al., 2015). Overload can result physical effects, for example with particles adhering to the surface of the organisms, causing impaired movement, or blockages of swimming appendages leading to increased energy cost for maintenance (Khan et al., 2017) . This requires an expert evaluation of the nanomaterial loading possibly combined with images of the biomass in the test system. A clear indication of testing artefacts caused by a high nanomaterial loading will negatively impact the reliability of the study.

17 - Is a sufficient number of replicates used? Is a sufficient number of organisms per replicate used for all controls and test concentrations? (Criticality level: ▲ ▲)

CRED: In general, the guideline requirements for number of replicates should be used. When a non-guideline 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 (….) lowers the reliability" (from Moermond et al. 2016).

nanoCRED: no nano-specific guidance for this criterion.

18 - Are appropriate statistical methods used? (Criticality level: ▲▲)

CRED: "In general, the guideline requirements for statistics should be followed and a description of the statistics is needed to assess the reliability of an end-point. 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" (from Moermond et al. 2016).

nanoCRED: no nano-specific guidance for this criterion.

19 - Is a concentration response curve observed? Is the response statistically significant? (▲/▲▲)

CRED: "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)" (from Moermond et al., 2016).

nanoCRED: Regarding concentration response curve: see explanation to criterion 13. There may be inherent characteristics on the nanomaterial test system which impedes the establishment of concentration response curves. However, whether the observed responses are statistically significant is considered to be important for study reliability. Hence, the criticality of this criterion depends on the specific type of test setup and output data. Expert judgement should be applied to evaluate the test results and the statistical significance of the observed response.

20 - Are sufficient data available to check the calculation of endpoints and (if applicable) validity criteria (e.g., control data, concentration–response curves)? (Criticality level: ▲ ▲ ▲)

CRED: "The availability of raw data is not a prerequisite for a study to be reliable. However, where "reliable with restrictions" (nR2) can be assigned if raw data are not reported, "reliable without restrictions" (nR1) 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" (from Moermond et al. 2016).

nanoCRED: no nano-specific guidance for this criterion.

Step 2: Assign reliability category

Based on the overall conclusion from the reliability evaluation the appropriate reliability category (as described in Table S2) is assigned to the study.

Table S2. Nano-specific reliability categories (adapted and further developed from Klimisch et al. (1997) and Moermond et al. (2016)). For information on "critical" (▲▲▲) and "important" (▲▲) reliability criteria see Table S1 and the main text.

Score	Description
nR1	**Reliable without restrictions**: All critical (▲▲▲) and important (▲▲) reliability criteria are fulfilled or partially fulfilled. The study is well designed, performed and documented. Nanomaterial properties and behaviour in the test system are extensively documented. The experiment has been carried out according to methods that are considered scientifically appropriate for ecotoxicity testing of nanomaterials and where the physico-chemical properties of the nanomaterial are considered in the test design. If (when) specific nanomaterial guidance or guidelines exist, the use of these may be considered favourable.
nR2	**Reliable with restrictions**: Most critical (▲▲▲) and important (▲▲) criteria are fulfilled or partially fulfilled. The study is generally well designed, performed and documented, but some minor flaws in the documentation or set-up may be present. Nanomaterial properties and behaviour in the test system is well documented. The experimental design and test method are considered scientifically appropriate for ecotoxicity testing of nanomaterials but may contain some minor flaws in documentation or setup.
nR3	**Not reliable:** Not all critical reliability criteria are fulfilled or partially fulfilled. This mainly concerns studies which have clear flaws in study design and/or how it was performed, and/or where the experimental design and test method are considered not to be scientifically appropriate for ecotoxicity testing of nanomaterials.
nR4	**Not assignable:** Information needed to make an assessment of the majority of the critical and important criteria is missing. This concerns studies or data from the literature which 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.

Step 3: Evaluate relevance

Relevance is evaluated using 13 relevance criteria (Table S3). The relevance criteria are based on the CRED evaluation method (Moermond et al. 2016) with nano-specific considerations for criterion number 2 and 4. The degree of fulfilment of each criterion is marked, either using "fulfilled", "partly fulfilled", or "not fulfilled", or by using a "fulfilment score", i.e. assigning numerical values (discrete values or a continuous range) on a scale going from 'the study did not address this criterion at all' to 'satisfactory fulfilment'.

If preferred, the risk assessor has the possibility to weight the criteria using the following descriptors: "not critical (▲), 'important' (▲▲) and 'critical" (▲▲▲). In addition, there is also a possibility to qualitatively summarize the evaluation. This can for example be done by multiplying a 'fulfilment score' with a 'weighing' score. If doing so, it is important to remember that a number can be a supplement to a qualitative summarising but it cannot replace the decision support that a detailed description of a study's strengths and weaknesses provides.

Table S3. *Relevance criteria adopted from the CRED evaluation method (Moermond et al., 2016) taking into account nano-specific concerns for criterion 2 and 4. See explanatory guidance text on how to interpret these criteria in the main text below.*

Number	Criterion
	General information
	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 sub-stance?

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?

Explanation of the relevance criteria (numbers refer to Table S1)

1 - Is the species tested relevant for the compartment under evaluation?

CRED: "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" (from Moermond et al. 2016).

nanoCRED: no nano-specific guidance for this criterion.

2 - Are the organisms tested relevant for the tested substance?

CRED: "Because the purpose of most assessments is to evaluate the potential risks of a substance to sensitive non-target 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 non-sensitive 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" (from Moermond et al. 2015).

nanoCRED: For studies on nanomaterials consideration may be given to the susceptibility of the test organism to particle exposure with regard to particle uptake and biological effects. For example, the normal feeding strategy of the organism may be considered (such as food size range, uptake mechanisms and location in water column).

3 - Are the reported endpoints appropriate for the regulatory purpose?

CRED: "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" (from Moermond et al. 2016).

nanoCRED: no nano-specific guidance for this criterion.

4 - Are the reported endpoints appropriate for the investigated effects or the mode of action of the test substance?

CRED: "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" (from Moermond et al. 2016).

nanoCRED: For nanomaterials with a known mode of effect, studies that investigate these effects may be considered more relevant. At present, modes of effect are still not well understood for nanomaterials although mechanisms linked to the production of reactive oxygen species has been proposed (Skjolding et al. 2016). Effects may also include uptake-related non-standard endpoints and be linked to the potential for translocation of the nanomaterial away from the point of initial uptake (such as the gut) into other parts of the organism. As a better scientific understanding of effects mechanisms develop, the way in which this criterion is evaluated may change.

5 - Is the effect relevant on a population level?

CRED: "Most frameworks consider only traditional test endpoints, such as mortality, growth, and reproduction, which are assumed to be linked to population sustainability. However, non-guideline tests may also report non-guideline 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" (from Moermond et al. 2016).

nanoCRED: no nano-specific guidance for this criterion.

6 - Is the magnitude of effect statistically significant and biologically relevant for the regulatory purpose (e.g., EC10, EC50)?

CRED: "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" (from Moermond et al. 2016).

nanoCRED: no nano-specific guidance for this criterion.

7 - Are appropriate life stages studied?

CRED: "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" (from Moermond et al. 2015).

nanoCRED: no nano-specific guidance for this criterion.

8 - Are the experimental conditions relevant for the tested species?

CRED: "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" (from Moermond et al. 2016).

nanoCRED: no nano-specific guidance for this criterion.

9 - Is the exposure duration relevant and appropriate for the studied endpoints and species?

CRED: "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" (from Moermond et al. 2016).

nanoCRED: no nano-specific guidance for this criterion.

10 - If recovery is studied, is this relevant for the framework for which the study is evaluated?

CRED: "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 " (from Moermond et al. 2016).

nanoCRED: no nano-specific guidance for this criterion.

11 - In case of a formulation, other mixture, salts, or transformation products, is the substance tested representative and relevant for the substance being assessed?

CRED: "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" (from Moermond et al. 2016).

nanoCRED: no nano-specific guidance for this criterion.

12 - Is the tested exposure scenario relevant for the substance?

CRED: "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" (from Moermond et al. 2016).

nanoCRED: no nano-specific guidance for this criterion.

13 - Is the tested exposure scenario relevant for the species?

CRED: "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" (from Moermond et al. 2016).

nanoCRED: no nano-specific guidance for this criterion.

Step 4: Assign relevance category

Based on the overall conclusion from the relevance evaluation the appropriate relevance category (as described in Table S4) is assigned to the study.

Table S4. Relevance categories (adopted from Moermond et al. 2016).

Score	Description
nC1	Relevant without restrictions: The study is relevant for the purpose for which it is evaluated.
nC2	Relevant with restrictions: The study has limited relevance for the purpose for which it is evaluated.
nC3	Not relevant: The study is not relevant for the purpose for which it is evaluated.
nC4	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 parameter

Step 5: Combine reliability and relevance

After evaluating both the reliability and relevance of a study its overall adequacy for regulatory purposes can be assessed. This can be done according to the approach summarised in Figure S1.
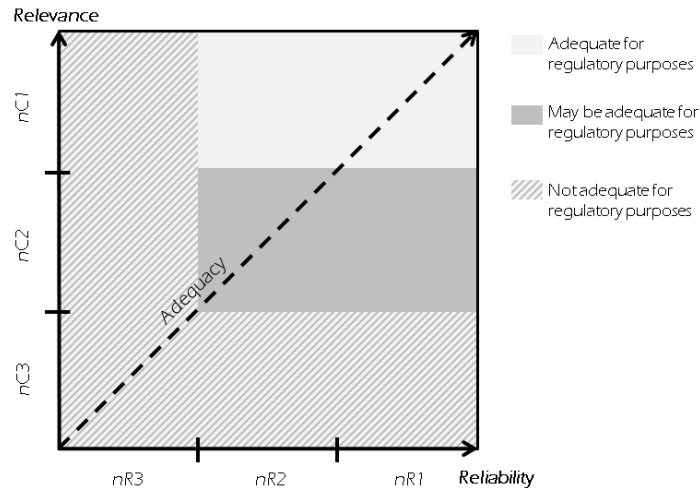

Figure S1. *Proposed two-dimensional approach to summarize the adequacy of ecotoxicity data in risk assessment of nanomaterials based on data reliability and data relevance. nR1-nR3 refers to the reliability categorisation of the study and nC1-nC3 to the relevance categorisation. Adapted from Ågerstrand et al. (2011) and Moermond et al. (2016).*

In this approach a study is considered 'adequate for regulatory purposes' if it is reliable without restrictions (nR1) OR reliable with restrictions (nR2) AND relevant without restrictions (C1). This type of studies would for example be used as key studies to derive PNEC values for risk assessment purposes. A study 'may be adequate for regulatory purposes' if it is reliable without restrictions (nR1) OR reliable with restrictions (nR2) AND relevant with restrictions (C2). This type of studies could for example be considered as supporting studies in a risk assessment – or could be considered for PNEC derivations in case limited data is available. Any study that is considered 'not relevant' (nR3) AND/OR 'not reliable' (C3) is considered 'not adequate for regulatory purposes'

References

Ågerstrand, Marlene, Magnus Breitholtz, and Christina Rudén. 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." Environmental Sciences Europe 23 (1): 17. doi:10.1186/2190-4715-23-17.

Baalousha, M. et al., 2016. The concentration-dependent behaviour of nanoparticles. *Environmental Chemistry*, 13(1), p.1.

Cupi, D., Hartmann, N.B. & Baun, A., 2016. Influence of pH and media composition on suspension stability of silver, zinc oxide, and titanium dioxide nanoparticles and immobilization of Daphnia magna under guideline testing conditions. *Ecotoxicology and Environmental Safety*, 127, pp.144–152.

Hartmann, N.B. et al., 2015. Techniques and Protocols for Dispersing Nanoparticle Powders in Aqueous Media—Is there a Rationale for Harmonization? *Journal of Toxicology and Environmental Health, Part B*, 18(6), pp.299–326.

Hartmann, N.B. et al., 2013. The challenges of testing metal and metal oxide nanoparticles in algal bioassays: titanium dioxide and gold nanoparticles as case studies. *Nanotoxicology*, 7(6), pp.1082–94. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22769854 [Accessed July 31, 2014].

Joint Research Center (JRC) 2016. URL: https://ec.europa.eu/jrc/en/scientific-tool/jrc-nanomaterials-repository (accessed 16th December 2016)

Klimisch, H.J., Andreae, M. & Tillmann, U., 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regulatory toxicology and pharmacology : RTP*, 25(1), pp.1–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9056496.

Moermond, C., Kase, R., Korkaric, M. & Ågerstrand, M. 2016. CRED: Criteria for reporting and evaluating ecotoxicity data. Environmental Toxicology and Chemistry, 35(5), pp.1297–1309.

NIST (2016). Standard Reference Materials. URL: https://www-s.nist.gov/srmors/ (accessed 16th December 2016)

OECD. 2012. Guidance on Sample Preparation and Dosimetry for the Safety Testing of Manufactured Nanomaterials. OECD Environment, Health and Safety Publications. Series on the Safety of Manufactured Nanomaterials.

ENV/JM/MONO(2012)/40.Petersen, E.J. et al., 2015. Adapting OECD Aquatic Toxicity Tests for Use with Manufactured Nanomaterials: Key Issues and Consensus Recommendations. *Environmental Science and Technology*, 49(16), pp.9532–9547.

Petersen, E.J. et al., 2014. Identification and Avoidance of Potential Artifacts and Misinterpretations in Nanomaterial Ecotoxicity Measurements. *Environmental Science & Technology*, 48(8), pp.4226–4246.

Petersen, E.J. & Henry, T.B., 2012. Methodological considerations for testing the ecotoxicity of carbon nanotubes and fullerenes: Review. *Environmental Toxicology and Chemistry*, 31(1), pp.60–72.

Skjolding, L.M. et al., 2016. A Critical Review of Aquatic Ecotoxicity Testing of Nanoparticles - The Quest for Disclosing Nanoparticle Effects. *Angewandte Chemie*, (November), pp.3–5.

Sørensen, S.N. & Baun, A., 2015. Controlling silver nanoparticle exposure in algal toxicity testing – A matter of timing. *Nanotoxicology*, 9(2), pp.201–209.

Sørensen, SN, Hjorth, R, Giron Delgado, C, Hartmann, NB & Baun, A 2015, 'Nanoparticle ecotoxicity—physical and/or chemical effects?' *Integrated Environmental Assessment and Management*, vol 11, no. 4, pp. 722-724., 10.1002/ieam.1683

van Hoecke, K., Quik, J.T., Mankiewicz-Boczek, J., Schamphelaere, K.A.D., Elsaesser, A., Meeren, P.V.D., Barnes, C., McKerr, G., Howard, C.V., Meent, D.V.D. and Rydzynski, K., 2009. Fate and effects of CeO2 nanoparticles in aquatic ecotoxicity tests. Environmental science & technology, 43(12), pp.4537-4546.