Reliability and relevance evaluation of ecotoxicity studies for use in risk assessment of chemicals.

-Working version-

Contents:
1. Definitions
2. Reliability evaluation
3. Relevance evaluation
4. References

1. Definitions
There are a number of definitions for relevance and reliability. We use the definitions that were also used by Klimisch et al. (1997) and which are taken up in the REACH guidance (REACH Guidance on information requirements and chemical safety assessment R.4):

**Reliability** - evaluating the inherent quality of a test report or publication relating to preferably standardised methodology and the way the experimental procedure and results are described to give evidence of the clarity and plausibility of the findings.

**Relevance** - covering the extent to which data and tests are appropriate for a particular hazard identification or risk characterisation.

From these definitions it can be deduced that reliability concerns the intrinsic scientific quality of the study (regardless from the purpose for which is it assessed), while relevance concerns the way the study can be used. Please note that a reliable study might be relevant in one framework but not in another. Also, a study with low reliability may still be relevant and could be used as supporting information.

**Example 1:** A reliable toxicity study with a sediment organism (e.g., *Lumbriculus*) is not relevant for the derivation of water quality standards, but may be relevant to derive sediment quality standards or to assess the environmental risk of a human medicinal product.

**Example 2:** An acute toxicity study with fish may be reliable, but not relevant to be used in the risk assessment of a hormone because the mode of action of hormones usually concerns reproduction and not toxicity to adult fish.

**Example 3:** A study with *Chironomus riparius* may not be reliable because the not-so-stable-compound was not measured while the rest of the study was reliable. However, it may be very relevant if high toxicity of the compound is shown in this study. If the concentration was measured, the endpoints would then probably have been even lower. Information like this, although coming from an unreliable study, may give an indication of strong toxicity to insects and can thus still be used to adjust the assessment factors in the risk assessment.
Weight of evidence /Plausibility is another issue regarding use of published data, but is for practical reasons not included in this ring test. For instance, a comparison with either historical test data or other test species is a weight of evidence approach thus these are not criteria in the current reliability and relevance assessment. If you want to be included in our work on weight of evidence, please indicate this on your questionnaire.

2. Reliability evaluation

Reliability is not only determined by the set-up, performance and evaluation of the experiment, but also by the reporting of the study. A properly reported study may be considered less or not reliable because of an inadequate test design (e.g. too few replicates), performance (e.g. control mortality too high) or data evaluation (e.g. inadequate statistics). Likewise, a study that was originally carried out in a scientifically sound way, may be classified less or not reliable if the description is very concise (e.g. experimental set-up is given as a reference to another report), or if various criteria that are considered important for interpretation of the test results cannot be checked (e.g. temperature data are not given) (Mensink et al., 2008).

The presented reliability criteria are divided into critical or non-critical criteria. Please note that some critical criteria are only critical for some compounds or organisms, e.g. for volatile compounds a closed system is essential, for algae the light conditions are essential, for some unstable compounds it is essential that concentrations are measured and a flow-through or renewal system is used, and for stable compounds a static system with nominal concentrations may be good enough. For specific details, see the description of the criteria below.

Another issue to keep in mind is that although the reliability concerns the intrinsic quality of the study, within one study one endpoint may be not reliable, while other endpoints are reliable. This could concern studies performed in different media (e.g., different pH’s and salinities). Endpoints like mortality and reproduction originating from the same study can also have a different reliability assessment if the study setup or the controls contains flaws for one endpoint but not for the other.

Example 4: A toxicity study with a compound to a freshwater fish in different media: distilled water, reconstituted water, and water with different salinities. The controls show high mortality in the tests performed in distilled water and high salinity, but no mortality in the tests performed in reconstituted water and low salinity. That means that the endpoints from the tests in distilled water and high salinity are not reliable because of the stress these test media caused. But: the endpoints from the tests in reconstituted water and low salinity, where the fish only experienced the stress from the compound under investigation, are reliable and relevant.

The reliability evaluation can be summarized into four different reliability scores (Table 1): Reliable without restrictions (Ri 1), Reliable with restrictions (Ri 2), Not reliable (Ri 3), and Not assignable (Ri 4) (table 1). In most frameworks, studies with reliability score 1 and 2 can be used in risk assessment or risk limit derivation, usually with equal weights. Studies with reliability score 3 and 4 cannot be used in the risk assessment, but are sometimes used as supporting information (depending on the
The relevance of the studies and the reasons for lower reliability. Please note that in general, if a study description lacks some details necessary to assign reliability, the authors can also be contacted for this information. (Please do not do this for the current ring test).

**Table 1.** Reliability scores (based on Klimisch et al., 1997).

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ri 1</td>
<td>Reliable without restrictions: All reliability criteria are fulfilled. (Studies or data from the literature or reports which were carried out or generated according to generally valid and/or (inter)nationally accepted testing guidelines or in which all parameters described are closely related/comparable to a guideline method.)</td>
</tr>
<tr>
<td>Ri 2</td>
<td>Reliable with restrictions: All critical reliability criteria are fulfilled. (Studies or data from the literature or reports in which the test parameters documented do not totally comply with specific testing guidelines, but are sufficient to accept the data or in which investigations are described which do not comply with a testing guideline, but which are nevertheless well documented and scientifically acceptable.)</td>
</tr>
<tr>
<td>Ri 3</td>
<td>Not reliable: Not all critical reliability criteria are fulfilled. (Studies or data from the literature or reports in which there were interferences between the measuring system and the test substance or which were carried out or generated according to a method which is not acceptable for the test compound or test organism, or which have other clear flaws in study setup, study conduction or reporting.)</td>
</tr>
<tr>
<td>Ri 4</td>
<td>Not assignable: Information needed to make an assessment of a critical criterion is missing. (Studies or data from the literature which do not give sufficient experimental details and which are only listed in short abstracts or secondary literature (books, reviews, etc.), or studies or reports of which the documentation is not sufficient for assessment of reliability.)</td>
</tr>
</tbody>
</table>

**Checklist for reliability criteria**

The reliability criteria are summarized in table 2. This checklist is a selection of the most important criteria and focuses on aquatic ecotoxicity studies. If more detailed criteria are deemed to be necessary to assess a publication, please refer to Ågerstrand et al. 2011, where a longer checklist of 62 criteria is available, or to Mensink et al., 2008.
Table 2. Reliability criteria. For a further explanation of the criteria, see the main text.

<table>
<thead>
<tr>
<th>Number</th>
<th>Criterion</th>
<th>Critical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>General information</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Before evaluating the test, check the physicochemical characteristics of your compound (handbooks/general sources). What is the solubility, $\log K_{ow}$, $pKa$, is the compound volatile, does it hydrolyse, photolyse, etc.?</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Is a description of endpoints and methodology available?</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td><strong>Protocol</strong></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Is a standard method (e.g., OECD/ISO) or modified standard used?</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Is the test performed under GLP conditions?</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>If applicable, are validity criteria fulfilled (e.g. control survival, growth)?</td>
<td>Yes. Criteria depend on test organism.</td>
</tr>
<tr>
<td>5</td>
<td>Are appropriate controls performed (e.g. solvent control, negative and positive control)?</td>
<td>Yes. Type of control depends on test substance and protocol.</td>
</tr>
<tr>
<td></td>
<td><strong>Test Compound</strong></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Is the tested substance identified clearly with name or CAS-number?</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>Are test results reported for the appropriate compound?</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Is the purity or the source reported? Is information on the formulation available (if appropriate)?</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td><strong>Test Organism</strong></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Are the organisms well described (e.g. scientific name, weight, length, growth, age/life stage, strain/clone)?</td>
<td>Yes. Necessary details depend on test organism.</td>
</tr>
<tr>
<td>10</td>
<td>Are the test organisms from a trustworthy source and acclimatized to test conditions? Have the organisms not been pre-exposed to test compound or other unintended stressors?</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td><strong>Exposure Conditions</strong></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Is the experimental system appropriate for the test substance and are appropriate test vessels used (e.g., static, flow-through, renewal; light/dark conditions; open/closed systems)?</td>
<td>Yes</td>
</tr>
<tr>
<td>12</td>
<td>Do the exposure concentrations not exceed water solubility? Or, if a solvent is used, is the solvent within the appropriate range and is a solvent control included?</td>
<td>Yes</td>
</tr>
<tr>
<td>13</td>
<td>Is a correct spacing between exposure concentrations applied? Is the exposure duration defined?</td>
<td>Yes</td>
</tr>
<tr>
<td>14</td>
<td>Have chemical analyses been performed to verify substance concentrations?</td>
<td>Yes. Exceptions possible.</td>
</tr>
<tr>
<td>15</td>
<td>Is the loading of the organisms within the appropriate range (&lt; 1 g/L)?</td>
<td>Yes, for hydrophobic compounds.</td>
</tr>
<tr>
<td></td>
<td><strong>Statistical Design and Biological Response</strong></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Is a sufficient number of replicates used? Is a sufficient number of organisms per replicate used for all controls and test concentrations?</td>
<td>Yes</td>
</tr>
<tr>
<td>17</td>
<td>Are appropriate statistical methods used?</td>
<td>Yes. Can also be recalculated by assessor afterwards if enough information is provided.</td>
</tr>
<tr>
<td>18</td>
<td>Is a dose response curve observed? Is the response statistically significant?</td>
<td>Yes. See at 17.</td>
</tr>
<tr>
<td>19</td>
<td>Are raw data available?</td>
<td>No. Not critical for reliability score 2, but essential for reliability score 1.</td>
</tr>
</tbody>
</table>
Explanation of the reliability criteria (criteria numbers from Table 2)
Knowledge on physicochemical parameters is essential to be able to decide on compliance with reliability criteria later on. Knowledge of the behavior of the compound in environmental media is important when deciding on the study setup. For instance, the solubility should not be exceeded and for volatile compounds other exposure conditions may be needed than for hydrophobic compounds. Many of these physico-chemical parameters can be found in chemical handbooks.

1 - Is a description of endpoints and methodology available?
Without a description of the endpoints and the methodology the study cannot be evaluated (Ri 4). Only a statement that the study was performed according to the OECD protocol is not enough; information on the test setup is necessary to decide if the test protocol is suitable for the compound and organisms tested.
The clarity of the description is not in itself a critical criterion for reliability. Although a clearly-written study report or published paper allows a fast and easy identification and use of necessary study data, this does not influence the quality of the research performed. Besides this, what is a ‘clear description’, depends very much on the reader. Some people prefer tables, others prefer text or figures. Some study reports or articles are written by non-native speakers, and may thus be harder to read than reports or articles from native speaker. In principle however, the language ability of the writer should not make a study more or less reliable. Thus, when a study is not clearly described it can still be reliable and this is not a critical criterion to decide on reliability. However, a well described study may enhance the feeling of the reader that the study is reliable.

2 - Is a standard method (e.g., OECD/ISO) or modified standard used?
Use of a standard method (OECD/ISO, US EPA or comparable) is not a critical criterion for a study to be reliable. Studies using non-standard methods may be equally reliable as guideline studies, as long as enough details on the setup are available to assess the scientific quality and reliability of the study. In addition, a guideline study may also be unreliable, if the study setup or results have flaws. For instance, when the choice of exposure conditions is not right for the compound under investigation, the control mortality is too high, or the validity criteria are not met, the study is unreliable despite the fact that it is a guideline study.

3 - Is the test performed under GLP conditions?
This criterion is not critical for the reliability of the study. GLP is a data quality system and requires adequate and permanent documentation of everything involved in the experimental test. GLP does however not reflect the quality of the study setup, the relevance of the endpoints, the reliability of the endpoints, and a good interpretation of the results. As such, for a study to be reliable, GLP could be recommended but should never be mandatory.

4 - If applicable, are validity criteria fulfilled (e.g. control survival, growth)?
In most test guidelines, validity criteria are mentioned to determine the validity of the test results. For instance, the OECD guideline 201 on algal toxicity, mentions that there should be exponential growth in the controls, and that the coefficient of variation should meet certain criteria. For the Daphnia acute toxicity study (OECD 202), validity criteria include control mortality and oxygen concentrations. If a non-guideline test is performed with standard species, validity criteria as described in the relevant guideline should be met. If non-standard species are used, expert
judgement is needed to assess whether the test organism resembles the standard test species enough to apply guideline validity criteria. For standard test species however, complying with guideline criteria for validity is critical for a study to be reliable.

5 - Are appropriate controls performed (e.g. solvent control, negative and positive control)?

Which controls to use depends on the test substance and/or the protocol applied. Next to the ‘normal’ negative controls (no solvent, no test substance), solvent controls needs to be tested in all cases where a solvent is used. The concentration of solvent in the solvent controls should be the highest solvent concentration used in the test systems, and the mortality in the solvent controls should not differ significantly from the non-solvent controls. If a solvent is used but no solvent control is tested, the solvent concentration in the controls is too low, or the mortality in the solvent control is higher than in the non-solvent controls, the study is not reliable (Ri 3).

In some cases a positive control (with a reference substance) is applied. The lack of a positive control does not decrease the reliability of a toxicity study. However, the presence of a positive control might give more confidence on the reliability of the study results.

If appropriate statistics are applied (see 18) the results of the negative control experiments themselves do not have to be reported for a study to be reliable (Ri 2). However, the conclusions on the solvent controls (if applicable) should be reported.

6 - Is the tested substance identified clearly with name or CAS-number? Are test results reported for the appropriate compound?

It is essential that it is known exactly which compound is tested. For instance, if a salt is tested, it should be known which salt is tested and how the results are expressed (as the salt or the base). The only exception for this is when the results are expressed in molarities instead of grams/litre, then it does not matter if results are expressed as the salt or the base. The lack of a CAS number does not decrease the reliability of the study, a CAS number can be easily retrieved on the internet.

Example 5: A test with Metformin hydrochloride is performed. It is not specifically mentioned if results are reported for the metformin base or the salt, but a NOEC and EC50 are tabulated for ‘metformin’. If actually metformin hydrochloride is meant, the toxicity values for metformin base would be lower. This decreases the reliability although results may still be used in risk assessment.

7 - Is the purity or the source reported? Is information on the formulation available (if appropriate)?

The purity of the substance should be reported or the source of the substance should be reported and be 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 any toxic effects by themselves and do not influence the toxicity of the substance of interest (EC, 2011). When the purity of the compound is < 90%, the nominal test results should be corrected for purity (EC, 2011).

Although the OECD guidelines only refer to a ‘suitable purity’, the reliability of a study which uses a low-purity substance should be lowered. When the actual test concentrations are measured however, this criterion becomes less important. If a formulation is used (e.g., when testing plant protection products), the other constituents of the formulation should be known, these other constituents should have no ecotoxicological effects, the amount of active substance in the formulation should be known, and it should be clear that the results are expressed in terms of active
substance. Also, the form of the formulation (e.g., microgranules) should not influence the actual uptake of the active substance by the test organisms.

8 - Are the organisms well described (e.g. scientific name, sex, weight, length, growth, age/life stage, strain/clone)?
When assessing reliability, it is essential to know which organisms have been used in the test. At least the name and some information on age or life stage should be known for a study to be reliable. Other information like weight, length, or strain/clone is not essential for the reliability assessment, but it may increases the confidence in the study. When examining hormonal substances, the sex of the organisms may influence the results. Thus, this information should then be known.

Example 6: 14-day fish toxicity studies with adult fish are usually regarded as acute toxicity studies. When tests are performed with fish larvae, however, a 14-days toxicity study is chronic since one or more sensitive life stages are included in the test (EC, 2011).

9 - Are the test organisms from a trustworthy source and acclimatized to test conditions? Have the organisms not been pre-exposed to test compound or other unintended stressors?
The source of the test organisms should be known and trusted. Test organisms should be acclimatized to the test conditions (e.g., water source, temperature) to avoid any unintentional stress due to these test conditions. The organisms used should also not be stressed by the test conditions applied or other (unintended) stressors and they should be healthy. When this stress is reflected in high control mortality, the study also becomes unreliable.
If test organisms could have been pre-exposed to the test compound (for instance, if field-collected), the results of the toxicity test can be biased because of community adaptation to toxic stress and therefore the study becomes unreliable.

10 - Is the experimental system appropriate for the test substance and are appropriate test vessels used (e.g., static, flow-through, renewal; light/dark conditions; open/closed systems)?
In most test guidelines a number of demands is described for the experimental system. Not all of these demands are mandatory; some depend on the substance and/or organism used. The demands of the appropriate test guidelines should be followed as closely as possible, also for non-standard test organisms. For instance, the test vessel should preferably be made of glass, but for some compounds this demand is more important than for others. For instance, when hydrophobic compounds are tested in paper cups or plastic containers the study becomes unreliable due to sorption of the compound to the container.
Static systems may be appropriate for short-term tests with stable compounds. However, for chronic tests static systems are usually not appropriate. For most compounds open systems may be used but for volatile compounds the test systems need to be closed for the study to be reliable.
Regular measurements of the test concentrations can confirm if the study setup has been appropriate: if the test concentrations were stable during the test, the study setup is reliable. If the 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 with restrictions. If not, the study is unreliable (Ri 3) or, if not enough details are provided, not assignable (Ri 4).
Example 7: Unfiltered natural water is used for an acute Daphnia toxicity study with a compound with log K_{ow} of 4.2; exposure concentrations are measured in unfiltered water. No information is provided on the amount of particulate matter in the water. Due to the high sorption of the compound it can be assumed that a significant amount of the compound has sorbed to the particulate matter. Assuming that the organisms are exposed to the dissolved concentration, in this experiment the actual exposure concentration is not known. Therefore, the study is unreliable. In contrast, when dissolved concentrations were measured and the total organic carbon was below 2 mg/L (OECD 202 requirement), the study would have been reliable.

11- Is the experimental system appropriate for the test organism; e.g., choice of medium or test water, feeding, water characteristics, temperature, light/dark conditions, pH, oxygen content?
The experimental system should be appropriate for the test organisms; for instance, freshwater species should not be tested in saltwater. However, what are appropriate test conditions depends very much on the organism tested and no specific guidance can be given here. For instance, algae need light to grow. But when testing a photodegradable substance, the experiment may be performed in the dark for fish and Daphnids. Temperature, pH and oxygen should preferably be stable and within the appropriate range for the organism and the compound. If there is a large variability among the controls or the control performance is in general not good, this may indicate that the test conditions were not appropriate and the study is not reliable.
Feeding is not allowed in acute toxicity studies due to interference with the test compound. However, for chronic studies feeding is often necessary to keep the animals alive. Feeding should then be very carefully, and all excess feed should be removed shortly after feeding. Sometimes aquatic organisms like Daphnia are tested in systems with sand or sediment. Endpoints from these studies can be regarded to be unreliable, especially when hydrophobic compounds are tested.

12 - Do the exposure concentrations not exceed water solubility? Or, if a solvent is used, is the solvent within the appropriate range and is a solvent control included?
In case very slightly soluble compounds are tested at concentrations up to the water solubility, and no ecotoxicological effects are observed, the test is in principle reliable. Depending on the uncertainty in the estimate of the water solubility, also test results that are above the estimated water solubility may be reliable. Expert judgement should be used to decide on this. Reports of precipitates may indicate that the solubility was exceeded. In this case, the test results become less reliable since actual concentrations do not equal nominal concentrations. Results from a test in which a slightly soluble compound was tested at nominal concentrations that are larger than 10 times the solubility, should by definition be regarded as unreliable (RI 3) (Mensink et al., 2008).
Use of a solvent may be used to prepare suitable concentrated stock solutions. Solvents may not be used to enhance the solubility in the test medium, and in any case the compounds used for this purpose should not be toxic to the tested species (EC, 2011). According to several OECD guidelines, the concentration of solvents should never exceed 0.01%. The inclusion of a solvent control is discussed under criterion 6.

13 - Is a correct spacing between exposure concentrations applied? Is the exposure duration defined?
When spacing between test concentrations is too large, the resulting endpoints are not reliable. Often a scaling factor of 3.2 is recommended. As a rule of thumb, a maximum scaling factor of 10 should be applied.

The exposure duration should also be defined for a study to be reliable. Sometimes, the exposure is very short (e.g., 1 day), while effects are observed during days or weeks after exposure. Results should then be expressed in terms of the exposure duration and not in terms of the duration of the whole experiment.

**Example 8:** A limit study is performed at two concentrations: 0.01 and 1 mg/L. At 0.01 mg/L no effect was observed and at 1 mg/L all organisms were dead. The NOEC could thus have been 0.01 mg/l but also 0.1 or 0.25 mg/l. Thus, from this study, no reliable NOEC, EC50 or LOEC can be derived because it is not known what the actual effect concentration is.

**14 - Have chemical analyses been performed to verify substance concentrations?**

It is important to know the actual exposure concentrations. Only acute static studies with stable compounds (information on stability should then be available from other experimental work or from the physical-chemical characteristics) may be reliable if no measurements are performed. In all other cases the exposure concentrations should be verified by analytical measurements. Depending on the compound and the test system, analysis of the concentration at the beginning of the test may be enough, but usually measurements should be performed at least at the beginning and the end of the test. There should be no major loss due to degradation, photolysis, volatilisation, hydrolysis, adsorption to glass, etc. At least at the start of the experiment, the test concentration should be close (80 – 120%) to nominal.

When results from the chemical analyses are reported, it should be clear if the concentrations reported are initial/final concentrations, mean or geometric mean concentrations, etc., and which of these concentrations are used to calculate the effect values.

The test design should normally be adequate to maintain >80% of the nominal concentration. If there is evidence that the concentration of the substance being tested has been satisfactorily maintained within ± 20 % of the nominal or measured initial concentration throughout the test, analysis of the results can be based on nominal or measured initial values and the study results are reliable.

If the loss of test substance is higher than 20 %, first it should be checked if the loss is caused by bad performance of the test (in which case the reliability is lowered), or by fast hydrolysis. The latter is done by inspecting the results of the hydrolysis test, and if these are not available (or in case of doubt), by consulting a specialist. Since high hydrolysis rate is an intrinsic property of the compound, it cannot be avoided. If all possible measures have been taken to maintain test concentrations (e.g. renewal; flow-through), and the test was thus performed in a technically adequate way, the test is considered reliable with restrictions. A > 20 % loss may thus be acceptable in some cases (Mensink et al., 2005). If the deviation from the nominal or measured initial concentration is not within the range of ± 20 %, analysis of the results should be based on geometric mean concentration during exposure or on models describing the decline of the concentration of the test substance (Mensink et al., 2008).

The method used to perform chemical analyses should be reported. However, if recovery efficiency of the method used and the limit of determination are reported, the reliability of the study increases. However, lack of information on recoveries and limit of detection does not make the study unreliable.
Conclusion: the need for measured concentrations depends on substance characteristics (e.g., if substance is volatile, instable or has adsorption potential). Measured concentrations at start should be within 80-120% of nominal. If test concentrations deviate more than 20% of nominal, it should be clear that all possible measures have been taken to avoid loss; test results should then be based on measured concentrations, preferably as time weighted average. In some cases, e.g. stable compounds in acute toxicity tests, nominal concentrations without measurements can be acceptable.

15 - Is the loading of the organisms within the appropriate range (< 1 g/l)?
Especially for hydrophobic compounds, organism loading should be taken into account to avoid loss of the test compound by sorption to biota. This is mainly relevant for studies with larger organisms, like fish and macrophytes. For instance, the OECD guideline for fish acute toxicity tests recommends a maximum loading of 1.0 g fish/litre for static and semi-static tests is; for flow-through systems higher loading can be accepted.

16 – Is a sufficient number of replicates used? Is a sufficient number of organisms per replicate used for all controls and test concentrations?
In general, the guideline requirements for the number of replicates should be used. When a non-standard study is evaluated, expert judgement needs to be used to assess if the design has been appropriate to obtain statistically reliable results. The use of pseudo replicates lowers the reliability.

17 - Are appropriate statistical methods used?
In general, the guideline requirements for statistics should be followed. When a non-standard study is evaluated, expert judgement is needed. A description of the statistics is needed to assess the reliability of the endpoint. A NOEC should be statistically significant and not just be determine by the eye.
If dose-response data are reported in a table or a graph, the assessor can use this data to calculate endpoints if these endpoints are not reported. For instance, when only an EC50 is reported but also a dose-response curve is available, an EC10 can usually be calculated from these data. Computer programs are available to translate graphs into individual datapoints (e.g., Techdig).

18 - Is a dose response curve observed? Is the response statistically significant?
If no dose-response curve is observed, no reliable endpoints can be derived and the study results should be regarded as unreliable. The only exception is when there is hysteresis; i.e., when at lower concentrations growth is increased and at higher concentrations a toxic effect is observed.
If no statistical significance is reported but a table or graph with dose response data is available, the assessor can calculate the endpoints and the statistics. If no standard errors are reported, calculating a statistically significant NOEC will not be possible. However, it may then be possible to calculate the EC10.
If the statistical method is provided but no dose-response graph or table is reported, the study can still be assigned to be reliable with restrictions.

19 - Are raw data available?
The availability of raw data on all treatments and replicates is not critical for the assignment of a reliability score 2. However, only when raw data are available the study can be assigned a reliability score of 1.

3. Relevance evaluation
Similar to the criteria for reliability, a distinction can be made between relevant and non-relevant studies (Table 3). As explained above, relevance is not in inherent quality of the study but also depends on the framework for which it is evaluated. It concerns the relationship between the test and the effect in the target species and whether the test method is meaningful and useful for a defined purpose (OECD, 2005b). Thus, the same study may be assigned different values for relevance when it is used for different purposes.

Similar to the reliability assessment, there are four relevance scores. However, to comply with phase I, and to avoid confusion, we have chosen to use only the relevance scores 1, 2, and 3 in the questionnaire. Other relevance scoring systems are also possible and we therefore propose different 2, 3 and 4 scoring systems in the questionnaire and ask you to select the one which is most practical to you.

Table 3. Relevance scores.

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Re 1</td>
<td><strong>Relevant:</strong> Studies or data from the literature or reports which are relevant for the purpose for which the study is evaluated.</td>
</tr>
<tr>
<td>Re 2</td>
<td><strong>Relevant with restrictions:</strong> Studies or data from the literature or reports which have a limited relevance for the purpose for which the study is evaluated.</td>
</tr>
<tr>
<td>Re 3</td>
<td><strong>Not relevant:</strong> Studies or data from the literature or reports which are not relevant for the purpose for which the study is evaluated.</td>
</tr>
<tr>
<td>Re 4</td>
<td><strong>Not assignable:</strong> Studies or data from the literature which do not give sufficient details and which are only listed in short abstracts or secondary literature (books, reviews, etc.), or studies or reports of which the documentation is not sufficient for assessment of relevance.</td>
</tr>
</tbody>
</table>

A study may be well conducted and fully reported, but the test endpoint may have little ecological significance (EC, 2011). In most frameworks, only test endpoints are used which can be linked to population sustainability, such as mortality and reproduction-related endpoints. In most guideline studies, unambiguous population relevant endpoints are reported. However, non-standard tests may also report non-standard endpoints which are very relevant, such as filtration rate and some behavioural endpoints. Non-relevant endpoints in most frameworks are endpoints like blood parameters, general behaviour, swimming speed, mRNA induction, in vitro tests, and coloration. Some relevance aspects can only be evaluated if the framework and the purpose for the risk assessment is known. For instance, a study (like for instance a bioaccumulation study) can be irrelevant for EQS derivation, but very relevant for PBT assessment, and vice versa. Sometimes a study contains useful toxicity information, but cannot be used directly for the purpose for which it is evaluated (e.g., risk limit derivation). In this case, it can still be used as supporting information. Examples are a NOEC value from a short-term test, or a value higher than the highest tested concentration or lower than the lowest tested concentration. The test can then still be classified as
reliable or reliable with restrictions, although the endpoints are not relevant for the specific purpose of evaluation (EC, 2011).

**Example 4:** A NOEC is available from an acute toxicity study with *Gammarus*. This NOEC is well below the lowest available EC50 values, and thus indicates that *Gammarus* is a very sensitive species. This information can then be used to adjust the assessment factor.

---

**Relevance criteria**
The relevance criteria are summarized in table 4. Each criterion is further explained in the text below.

<table>
<thead>
<tr>
<th>Table 4. Checklist for relevance criteria.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number</strong></td>
</tr>
<tr>
<td><strong>General</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td><strong>Exposure relevance</strong></td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
</tbody>
</table>

**Explanation of the relevance criteria (Criteria numbers from table 4)**
Before evaluating the test for relevance, check why you are evaluating this study. The relevance of the study might be different for different purposes (e.g., EQS derivation, PBT assessment, dossier evaluation for marketing authorisation), also depending on the framework for which the evaluation is requested. In contrast to reliability, which is an intrinsic property of the study (and should not differ among frameworks), the relevance depends strongly on the framework for which a study is evaluated.
1 – Are the organisms tested relevant for the aquatic compartment and the tested compound?
The species tested should be relevant for the compartment under evaluation. In this case, because we focus on aquatic toxicity studies, the species should be relevant for the aquatic compartment. For instance, soil organisms like nematodes, even when tested in aqueous medium, are not relevant for the aquatic compartment. When endocrine disrupting compounds are tested, effects might be different for males or females. Thus, a distinction should then be made between male and female results and the relevance of the study results should be assessed for both sexes. For a study to be relevant, the test organisms do not necessarily have to be a validated test species. Also tests with other species can be very relevant.

2- Are the reported endpoints appropriate for the investigated effects or the mode of action?
Usually, standard endpoints like mortality and reproduction are assessed. Non-standard endpoints may however also be very valuable for risk assessment or risk limit derivation. When a risk assessment is performed for compounds with a specific mode of action, studies which assess this mode of action are most relevant. For example, fish biomarkers, vitellogenin (VTG), secondary sex characteristics (SSC) and sex ratio are considered to indicate endocrine disrupting chemicals interfering with estrogens, androgens and steroidogenesis pathways (OECD, 2012). These biomarkers are not useful for indicating other modes of action like the glucocorticoid receptor pathway.

3- Is the effect population relevant?
Most frameworks use only population-relevant effects in the risk assessment or risk limit derivation. In most guideline studies, unambiguous population relevant endpoints are reported. However, non-standard tests may also report non-standard endpoints which are very relevant, such as filtration rate and some behavioural endpoints. Non-population relevant endpoints in most frameworks are endpoints like blood parameters, general behaviour, swimming speed, mRNA induction, *in vitro* tests, and coloration.

4 - Is the magnitude of effect (e.g. EC10, EC50) relevant according to the guideline?
For the derivation of chronic risk limits, usually EC10 and NOEC value are used. EC50 values derived from chronic toxicity tests are usually not relevant. Likewise, for the derivation of acute risk limits, EC50 values are used and NOECs or EC10s derived from acute studies are less relevant.

5 - Are appropriate life-stages studied?
The life-stage studies should fit with the experimental design and the purpose of the study. For instance, an ELS test with fish larvae is very relevant for compounds that affect development, but not relevant for compounds that affect reproduction or male/female ratios. For those compounds, other tests may be more relevant.

6 - Are the experimental conditions relevant for the tested species?
Usually, aquatic organisms are tested in water and aquatic tests should not be performed with soil organism. Although this seems obvious, this criterion is not always met. Freshwater species should be tested in freshwater and saltwater species in saltwater. If this is not the case, the result is not relevant. Moreover, the route of administration should reflect the compartment under investigation.
For sediment organisms, sediment needs to be present in the test system. For instance, a study with the sediment-organism *Lumbriculus* in only-water systems is not relevant.

**7- Is the time of exposure relevant and appropriate for the studied endpoints and species?**
The time of exposure should be in line with the endpoints studied. Depending on the species, the exposure time may vary. For algae, usually the maximum exposure time is 4 days, but depending on the species studies with 7 days may also be used. Although most guidelines recommends 96 hours for a fish acute toxicity tests, this does not mean that a 5-day or 10-day test is not relevant. When studying chronic effects, the exposure time should be long enough to include the most sensitive life-stage or a whole life-cycle (See also example 2).

**8- If recovery is studied, is this relevant for the framework for which the study is evaluated?**
In most frameworks, recovery is not taken into account. Endpoints which include recovery are thus not relevant in these frameworks. In the plant protection product (PPP)-framework however, recovery may be taken into account and endpoints including recovery are very relevant.

**9- Is the substance tested representative and relevant for the substance being assessed?**
A substance may be tested as a pure active substance or in a formulation. For EQS derivation, tests with the formulation may be less relevant while in the PPP framework they may be very relevant. For unstable compounds, it should be known if metabolites are formed and if these metabolites are also toxic. If the substance causing the effect is not the substance under investigation, the results are unreliable.

**10 - Is the tested exposure scenario relevant for the substance?**
In some cases, the exposure scenario is not relevant for the substance tested. For instance, a >100 day chronic fish study may not be relevant for a compound that degrades within a day. For a realistic risk assessment, the exposure scenario would then be much more suitable for possible metabolites. In contrast, for compounds that are stable, longer exposure durations may be relevant.

**11- Do the tested concentrations relate to measured or predicted environmental concentrations (if available)?**
Testing the compound at concentrations close to the measured or predicted environmental concentrations, may increase the credibility of the total risk assessment. However, depending on the sensitivity of the species, environmental concentrations may be not relevant at all in toxicity testing. A non-sensitive species may have a NOEC far above the environmental concentrations, while a sensitive species may have 100% effect at environmental concentrations and a NOEC far below. Both these tests are relevant, especially when a species sensitivity distribution can be calculated from all available data.
4. References


Please help improve this guidance document. If you have specific comments which are not included in the questionnaire, or if you have additional examples, please contact:
caroline.moermond@rivm.nl