SciRAP criteria and guidance for assessing methodological quality of *in vivo* toxicity studies

The guidance is meant to aid in the evaluation of methodological quality, i.e. design and conduct, of *in vivo* toxicity studies according to the SciRAP criteria (available online at www.scirap.org). Note that the guidelines are not intended to aid in the interpretation of study results. The SciRAP guidance has been primarily based on recommendations and examples provided in OECD test guidelines and corresponding guidance documents. Changes and updates may become necessary as scientific knowledge advances. This guidance may not cover all relevant aspects of each criterion for all study types. For more details, the evaluator is referred to other guidance for the design, conduct and interpretation of toxicity testing, e.g. OECD test guidelines and guidance documents. We also welcome comments that can help improve the SciRAP guidelines.

Criteria	**Guidance**
Test compound and controls	
1. The test compound or mixture was unlikely to contain any impurities that may significantly have affected its toxicity.	Purity of the test compound, or the composition of substances in a mixture can potentially affect study results. Purity and composition is also an important aspect to consider in terms of the relevance of the test compound to the compound being risk assessed. Ideally, in the case of single compounds, the test chemical should be of the highest available purity. Significant impurities, or isomers of the test compound, are more likely to be present, and/or to impact toxicity for certain compounds. For example, PCBs (individual or in mixtures) are often contaminated with low levels of potentially highly toxic dioxins. The measured toxicity of the test compound may then be due to the contaminant. In such cases information about the level of purity and composition is critical. <u>How to judge this criterion:</u> Fulfilled – The test compound has been clearly identified and characterized and is of sufficient purity. In cases of mixtures, the composition of substances is well characterized and their individual purities are sufficient. Partially fulfilled – The purity of the test compound has not been described but it is unlikely that impurities are present that would significantly affect the results of the study.

	Not fulfilled – The test compound or mixture is likely to contain impurities that can affect study results.
2. An appropriate vehicle was used that is not expected to interfere with the absorption, distribution, metabolism, excretion or toxicity of the test compound.	A vehicle is "any agent which serves as a carrier used to mix, disperse, or solubilize the test item or reference item to facilitate the administration/application to the test system" (OECD 1998). The choice of vehicle will be determined by the solubility of the test compound as well as the route of administration used. It should be noted that both the vehicle and the route of administration may significantly affect the toxicokinetics and metabolism of the test compound. In cases where the compound is administered dermally it is also important to consider the influence of the vehicle on how the compound may penetrate the skin. In regard to the choice of vehicle, current OECD test guidelines for oral administration recommend that "the use of an aqueous solution/suspension be considered first, followed by consideration of a solution/emulsion in oil (e.g. corn oil) and then by possible solution in other vehicles" (OECD 2015). Ideally, the toxic characteristics of the vehicle, as well as the stability and homogeneity of the test chemical in the vehicle should be reported. If information concerning the *potential toxicity* of the vehicle is missing from the study it should not automatically lead to the judgment "not fulfilled" or "cannot be determined" for this criterion. The approach to evaluation may be to consider if the vehicle is clearly inappropriate, e.g. potentially toxic or affecting skin permeability or toxicokinetics. If information about which vehicle was used is completely missing the criterion should be judged as "cannot be determined" and a comment motivating this judgment should be included. <u>How to judge this criterion:</u> Fulfilled – water or another common and historically well characterized vehicle was used, the vehicle was appropriate considering the solubility of the test compound, and there are no other aspects that raise concern. Partially fulfilled – the vehicle was not well characterized or is not commonly used in this context but there are no obvious concerns that it interferes with the absorption, distribution, metabolism, excretion or toxicity of the test compound.

	Not fulfilled – the vehicle used is inappropriate considering the solubility of the test compound or otherwise, or is clearly toxic.
3. A concurrent negative control group was included.	A concurrent negative control group should always be included as it is critical for determining treatment-related effects. The negative control group can be either untreated or vehicle-treated. However, in studies where a vehicle is used to administer the test compound it is critical that a vehicle-treated control group is included. In certain cases, it may be useful to also include a completely untreated group for identification of any influence on results from the vehicle. Control animals should be handled in the same way as treated animals. It is also important that animals in the control and treatment groups are the same age since some toxicological effects are age-dependent, e.g. may represent acceleration and/or enhancement of age-related changes. Historical control data from the same laboratory using the same methods and relating to animals of the same strain, age and sex, and supplier, as those used in the study may be very useful. However, such data should not provide the only negative control data for statistical analyses as biological parameters in laboratory animals can vary significantly over time. Therefore, if a study includes only historical negative control data this criterion should be judged as "not fulfilled". <u>How to judge this criterion:</u> Fulfilled – a concurrent negative control group was included. Not fulfilled – no negative control was included or only a historical negative control was referred to.
Animal model and housing conditions	
4. A reliable and sensitive animal model was used for investigating the test compound and selected endpoints.	The choice of animal model (test species, strain, sex, etc.) is based on a number of considerations, including knowledge regarding species differences in terms of pharmacology, repeat-dose toxicology, metabolism, toxicokinetics and route of administration. Rodents (rats or mice) are commonly recommended for *in vivo* testing in current OECD test guidelines and are well characterized in terms of the reliability and sensitivity, as well as relevance to humans of different biological parameters and endpoints. Thus, it is specifically important that the study authors have justified their choice of animal model if other species have been used. It should be noted that, for investigation of certain endpoints,

	other species may be more sensitive and preferable. For example, rabbits are commonly recommended for teratology studies (OECD 2008). Similarly, available information about species differences in the toxicokinetics of a compound may warrant testing in a specific species. The evaluator is referred to regulatory test guidelines (e.g. OECD or US EPA) for discussions of the most appropriate test species for different study types. Reliability, in this context, refers to whether the animal model has been shown to generate reproducible results for the type of endpoints investigated. The sensitivity of the animal model relates to the ability to detect changes in the endpoints investigated in the model. E.g. different strains of rats may exhibit differences in the sensitivity to the effects of estrogenic compounds <u>How to judge this criterion:</u> Fulfilled – The animal model used is not suspected to be insensitive or unreliable. Not fulfilled – there is available information that indicates that the animal model is either insensitive or clearly unreliable for studying the test compound or for investigating the endpoints considered. Or the expected outcome is lacking from concurrent positive controls, if included, indicating that the test methods or animal model is insensitive.
5. Animals were individually identified.	In order to ensure reliable administration of the test compound, allocation to treatment groups and different tests, as well as recording of observations and test results, it is important that animals are individually identified. <u>How to judge this criterion:</u> Fulfilled – it is stated that animals were individually identified, the specific method for identification does not have to be described. Partially fulfilled – it is not clearly stated whether or not animals were individually identified, but it may be inferred from other information reported for the study design and conduct

	Not fulfilled – it is stated that animals were not individually identified, or this can be inferred from other information reported for the study design and conduct.
6. Housing conditions (temperature, relative humidity, light-dark cycle) were appropriate for the study type and animal model.	Housing conditions and handling may influence animal behavior and physiological response to stress and, consequently, study results. Importantly, variability in housing conditions may lead to increased variability in results and decreased sensitivity of the tests conducted. Different housing conditions apply to different species and different types of studies. Descriptions of standard conditions may for example be found in OECD test guidelines relevant to different types of studies and in corresponding guidance documents (http://www.oecd.org/env/ehs/testing/oecdguidelinesforthetestingofchemicals.htm). Guidance is also provided in the US National Research Council's "Guide for the Care and Use of Laboratory Animals" (https://grants.nih.gov/grants/olaw/Guide-for-the-Care-and-use-of-laboratory-animals.pdf) Housing conditions are often incompletely reported in studies published in the peer-reviewed literature, therefore it might be useful to keep in mind that this criterion may often be judged as partially fulfilled for such studies, and the impact of lack of reporting on total study reliability should be carefully considered. If no housing conditions were reported this criterion should be judged as "cannot determine" and a comment justifying the judgment should be made. <u>How to judge this criterion:</u> Fulfilled – housing conditions have been fully described and were in line with standard recommendations relevant to the study type and animal model. Partially fulfilled – some of the housing conditions were in line with standard recommendations relevant to the study type and animal model. Others deviated from standard recommendations or were not reported. Not fulfilled – all housing conditions deviated from standard recommendations relevant to the study type and animal model.
7. The number of animals per sex in each cage were	The number of animals housed together may have an effect on behavior and other biological parameters. Generally, laboratory animals should be housed in pairs or groups, unless the species is

appropriate for the study type and animal model.	naturally solitary. Crowding should also be avoided as it induces stress that affects e.g. hormone levels and development. Scientific and practical aspects connected to the type of study influence how animals are housed together. Recommendations and requirements for the number of animals per cage relevant for different study types can be found in OECD test guidelines and corresponding guidance documents. Single housing may be recommended in some cases, e.g. in acute toxicity tests and in inhalation studies using aerosol exposure. Individual housing may also be necessary e.g. for pregnant dams and for males after mating, as well as during certain procedures, such as the use of metabolism cages. According to general guidelines for laboratory animal science, single housing should be restricted to the shortest time possible. Standardization of litter size by culling is sometimes conducted. Descriptions and recommendations for this procedure are provided in OECD test guidelines for developmental toxicity studies. <u>How to judge this criterion:</u> Fulfilled – the number of animals per sex and cage were in line with standard recommendations relevant to the study type and animal model. Partially fulfilled – the number of animals per cage deviated somewhat from standard recommendations relevant to the study type and animal model, however scientific and/or practical justifications for these deviations were provided. Not fulfilled – the number of animals per cage deviated significantly from standard recommendations relevant to the study type and animal model and no scientific or practical justification was provided.
8. The test system is unlikely to contain contaminants that could affect study results, such as organic pollutants, pesticide residues, heavy	Materials used in cages, water bottles and any physical enrichment should be considered, e.g. in terms of releasing substances that may affect study results. It should be ensured as far as possible that feed and drinking water are free from contaminants, such as pesticide residues, persistent organic pollutants, heavy metals and mycotoxins, as well as phytoestrogens. Phytoestrogen content is specifically critical in studies where endocrine activity/disruption is being investigated. For guidance on appropriate phytoestrogen levels in feed see

metals, and mycotoxins, as well as phytoestrogens.	e.g. OECD TG 440 (OECD 2007b). Ideally, feed and water should be tested for the presence of contaminants and phytoestrogens. Similarly, the bedding material should be considered, especially if endocrine activity/disruption is being investigated, since it may contain naturally occurring estrogenic or antiestrogenic substances. E.g. corn cob appears to be antiestrogenic and affects cyclicity in rats (OECD 2007b). Specifically, phytoestrogen content should be minimized in the bedding material in these cases. A full report of possible contaminants is seldom provided in studies published in the peer-reviewed literature, therefore it might be useful to keep in mind that this criterion may often be judged as partially fulfilled for such studies, and the impact of lack of reporting on total study reliability should be carefully considered. <u>How to judge this criterion:</u> Fulfilled – no contaminants that could have influenced study results are suspected and/or feed, water, bedding and other materials have been analyzed and controlled for relevant contaminants. Partially fulfilled – some contaminants have been controlled for or analyzed but there may potentially be other contaminants present. Not fulfilled – it is likely that the test system was contaminated in a way that could affect study results, e.g. a bedding material known to contain estrogenic or antiestrogenic substances was used in a study investigating endocrine endpoints.
Dosing and administration of test compound	**Although the aspects evaluated in this section are linked to the relevance (sensitivity) of the test method they are considered important factors that have bearing also on the validity, accurateness and robustness of a toxicity study and are therefore included here in the evaluation of reliability.**
9. The allocation of animals to different treatments was randomized.	Animals should be randomly assigned to control and treatment groups. Randomization is applied as one measure to avoid that statistically significant results arise by chance alone. <u>How to judge this criterion:</u>

	Fulfilled – animals were randomly assigned to different treatment groups using an appropriate method Partially fulfilled – not clearly reported if animals were randomized but it can be inferred from other information reported for the study design and conduct. Not fulfilled – animals were not randomly assigned to treatment groups
10. The route of administration was appropriate and not likely to interfere with the study results.	*NOTE: The relevance of the administration route for human exposure scenarios and health risk assessment should not be specifically considered here. Application of this criterion for the evaluation of study reliability should consider any influence of the administration route on the validity, accurateness and robustness of the study results.* In general, for repeated dose and long term/chronic studies it is recommended that the test compound is administered orally, by dietary admixture or in drinking water, by gavage or in capsules (for non-rodents). If another route of exposure was used in such a study, the scientific and/or practical reasons for this should generally be justified by the study authors. Gavage is often used as it allows for delivery of a precise and consistent dose. Administration in feed or drinking water may be chosen as it is less invasive than gavage and sometimes better mimics human exposure scenarios (i.e. is more relevant). It is important to note that the toxicokinetics of the test compound may vary, and consequent toxicity may be different, if administered via gavage as compared to administration in feed or water. If administration is via food or water in perinatal studies offspring may be exposed both indirectly via maternal milk (if the test compound is transferred) and directly from feed/water in the last part of the lactation period, as they gradually start to consume food and water (from around postnatal day 14 in rats) and they may actually be consuming a higher dose per kg/bw than the adults. (OECD 2008) Dermal administration and inhalation may often pose additional practical and technical difficulties compared to oral administration or injection. For this reason, dermal administration is specifically not recommended for reproductive toxicity studies according to OECD test guidelines and guidance documents (OECD 2008). In terms of inhalation, many factors can affect e.g. deposition and retention of the test compound in the respiratory tract and the dose is dependent on how much of the compound is delivered to the exposure chamber in a respirable form. (OECD 2002b). It should be noted that if pups are exposed in inhalation chambers they may receive inhaled and dermal doses simultaneously, depending on the equipment (OECD 2008).

	Importantly, in reproductive/developmental studies, dams and pups should not be separated for long periods of time (several hours) to allow for exposure of one or the other (e.g. inhalation studies). (OECD 2008) How to judge this criterion: Fulfilled – the recommended route of administration was used and is considered to not influence study results in this case. In case an alternative administration route was used this was explicitly and correctly justified by the study authors. Partially fulfilled – another administration route than recommended in test guidelines was used and was not justified by the study authors. However, the chosen administration route is unlikely to have influenced study results. Not fulfilled - an alternative administration route was used and is suspected to have influenced study results.
11. The timing and duration of administration were appropriate for investigating the included endpoints.	OECD test guidelines and corresponding guidance provide recommendations for timing and duration of administration of the test compound for different types of studies. In general, the dosing regimen should "maximise the sensitivity of the test without significantly altering the accuracy and interpretability of the biological data obtained" (OECD 2002b). Timing and duration should be considered specifically in terms of covering sensitive periods of development (e.g. "period of male sexual differentiation in late gestation" (OECD 2008)). In certain cases, it is also relevant to consider timing of administration in relation to when measurements of toxicological outcomes are conducted. For example, when investigating effects on behavior the potential of the administration to produce acute effects on behavioral measures should be considered, especially where the test substance is administered directly to offspring daily (OECD 2008). How to judge this criterion:

	Fulfilled – the timing and duration of administration of the test compound is in line with general recommendations for the study type, is not likely to interfere with the measurements conducted, and cover sensitive periods of development, where relevant. Partially fulfilled – the timing and duration of administration of the test compound deviates somewhat from standard recommendations, however a scientific or practical justification is provided and sensitive periods of development are covered. Not fulfilled - the timing and duration of administration of the test compound is significantly different from general recommendations for the study type without being justified, and/or is likely to directly interfere with toxicological outcomes/measurements, and/or do not cover sensitive periods of development, where relevant.
12. The frequency of administration is appropriate for investigating the included endpoints.	The appropriate frequency of administration depends on the dose, physicochemical properties and toxicokinetics of the test compound, as well as practical considerations. "Responses produced by chemicals in humans and experimental animals may differ according to the quantity of the substance received and the duration and frequency of exposure, e.g. responses to acute exposures (a single exposure or multiple exposures occurring within twenty-four hours or less) may be different from those produced by subchronic and chronic exposures." (OECD 2002b) Fulfilled – the frequency of administration of the test compound is in line with general recommendations for the study type and considering the inherent properties of the test compound. Or, if not in line with general recommendations, adjustments have been justified by the study authors. Not fulfilled - the frequency of administration of the test compound is significantly different from general recommendations for the study type without being justified, or seems inappropriate considering the inherent properties of the test compound.
Data collection and analysis	
13. The allocation of animals to different tests and	Animals should be randomly assigned to different tests and/or measurements. Randomization is applied as one measure to avoid that statistically significant results arise by chance alone.

measurements was randomized.	How to judge this criterion: Fulfilled – animals were randomly assigned to different tests and measurements using an appropriate method Partially fulfilled – not clearly reported if animals were randomized but it can be inferred from other information reported for the study design and conduct. Not fulfilled – animals were not randomly assigned to tests and measurements.
14. Reliable, and sensitive test methods were used for investigating the selected endpoints.	The reliability of the methods refers to whether they are known to generate reproducible results for the type of endpoints investigated, e.g. if the methods have been validated across different laboratories. The sensitivity of the methods relates to the ability to detect changes in the endpoints investigated. Studies conducted according to standardized and validated test guidelines (such as OECD test guidelines) are often considered to be reliable and adequate for risk assessment. However, it is important to keep in mind that adherence to standardized test guidelines does not automatically ensure the sensitivity of the methods applied. Further, sensitivity of the methods may in some cases be influenced by how the protocols are utilized (OECD 2008). How to judge this criterion: Fulfilled – there is no information that suggests that the test methods are insensitive or unreliable in this context. Partially fulfilled – it is suspected that one or more of the methods applied may be insensitive or unreliable. Not fulfilled – there is available information that indicates that one or more of the methods applied is either insensitive or clearly unreliable for studies of the test compound or for investigating the endpoints considered. Or the expected outcome is lacking from concurrent positive controls, if included, indicating that the methods or animal model is insensitive.

15. Measurements were collected at suitable time points in order to generate sensitive, valid and reliable data.	This criterion covers several aspects concerning the timing of measurements and collection of data. Overall, to avoid introducing potential bias and to generate robust data, it is most important that tests and/or measurements are performed under the same conditions in all treatment groups. 1. Data should be collected at the correct time point in relation to the time needed to detect treatment related effects. In regard to specific developmental effects, these may only become apparent at a certain age, relating e.g. to behavioral ontogeny or onset of puberty. In addition, the time point for measurements and data collection should be chosen to avoid influence from any acute effects of the test substance administration (OECD 2008). OECD test guidelines provide recommendations for the timing of measurements and data collection in different study types. 2. Data should be collected at the same age across treated and control animals. For some developmental effects in rats and mice investigated during pregnancy or early after birth the time of day when measurements are performed is critical since development is rapid and differences between controls and treated animals may otherwise only represent differences related to (gestational) age (OECD 2008). 3. Data should be collected so that the time of day does not influence measurements. For example, responses in behavioral testing in nocturnal animals like mice and rats is likely to produce different behavior during the day than during the night. For such reasons reversed lighting conditions may be applied to test nocturnal animals during the day. How to judge this criterion: Fulfilled – The timing of tests and measurements were appropriate to detect sensitive effects and there are no related aspects that are likely to influence the reliability of the results. Conditions have been the same for treated and control animals. Partially fulfilled – Some, but not all, aspects of timing were appropriate. Importantly, there are no critical issues that raise concern, e.g. that control and treated animals were tested at different age/time points. Not fulfilled - The timing of tests and measurements were not appropriate. E.g. it is likely that sensitive treatment related effects have been missed, or there are other aspects that are likely to have influenced the reliability of the results. And/Or control and treated animals were tested at different age/time points.

16. A sufficient number of animals per dose group were subjected to separate tests/data collection/measurements to generate reliable and valid results.	Sample size should be large enough to ensure sufficient statistical power to detect any effects in the endpoints measured. This includes considerations of the background incidence and variability of the measured effects, as well as the method of analysis. Excessive losses of animals in treatment groups that could affect statistical power should be noted. OECD test guidelines provide recommendations for number of animals per treatment group for different study types and endpoint measurements. However, primary consideration should be given to justifications for sample size provided by study authors, if stated. <u>How to judge this criterion:</u> Fulfilled – a sufficient number of animals was included in the different treatment groups and loss of animals during the study is not likely to have substantially affected statistical power. Partially fulfilled – a lower than usual number of animals was used, which may have caused lower sensitivity/statistical power of the study. Not fulfilled – the number of animals in each treatment group was clearly insufficient or there was substantial loss of animals during the study that may have affected statistical power.
17. The statistical methods have been clearly described and do not seem inappropriate, unusual or unfamiliar.	The choice of statistical analyses will depend on the type of study and the nature of the endpoints measured. OECD test guidelines and corresponding guidance documents provide some recommendations for statistical tests (e.g. OECD 2002a,b) as well as for considerations to be made in statistical analyses of different types of tests. Evaluation of this criterion also includes considering if the correct statistical unit was used. For example, it is generally recommended that the litter (or dam) is the statistical unit in developmental toxicity studies to account for litter effects. Correlations across litter mates due to genetic and/or prenatal conditions can have considerable influence on the statistical significance of results (e.g. Holson et al. 2008; Li et al. 2008). To control for litter effects, either only one pup per sex and litter is submitted to each test/measurement in the study, or all pups are examined and litter effects are accounted for in the statistical analyses. For certain endpoints, e.g. malformations, it might be

	warranted to examine all pups as it increases the statistical power and not all pups are identical. Similarly, examining many pups per litter greatly enhances the ability to detect low dose effects (OECD 2008). The size of litter effect varies depending on endpoint measured, dose (being larger at high dose levels), and chemical mode of action. In general, normality of the data should have been checked and the choice of parametric or non-parametric tests should have been based upon that result. <u>How to judge this criterion:</u> Fulfilled – the statistical methods have been clearly described and do not seem inappropriate, unusual or unfamiliar. Partially fulfilled – unusual or unfamiliar methods were applied in the statistical analyses but do not seem clearly inappropriate. Not fulfilled – no statistical tests were used, or the tests used are clearly inappropriate for the study type and/or endpoints measured.
18. Are there any other aspects of study design, performance or reporting that influence reliability?	In this section any additional factors of the study design or conduct that are not covered by the criteria above and that the evaluator considers may increase or decrease reliability of the study should be considered. These may vary on a case-by-case basis and can for example include, but are not limited to, factors such as: • If the test method has been validated, e.g. by assessment of repeatability within a laboratory and reproducibility of the method at multiple laboratory sites. • If blinding of experimenters to the allocation to treatment groups during testing and analyses of data was applied. Application of blinding may be hampered by practical problems but testing and analyses without knowledge of treatment group may eliminate some bias that could influence results, especially in behavioral studies.

	• If internal dose was measured. • If testing in behavioral studies was conducted using automated methods. • If measurements of specific endpoints were collected by the same or different observers. • If there was any unusual or unexplained mortality or premature loss of test animals. • If declarations of conflict of interest or sources of funding raise concern of possible bias or are missing. Note: this criterion is not included in the color profile read-out for the study generated in the SciRAP tool. Comments made here will be shown in the excel file with the color profile for the study and should be considered in parallel with the color profile when conducting the categorization of study reliability.

References

Holson RR, Freshwater L, Maurissen JP, Moser VC, Phang W. 2008. Statistical issues and techniques appropriate for developmental neurotoxicity testing: A report from the ilsi research foundation/risk science institute expert working group on neurodevelopmental endpoints. Neurotoxicol Teratol 30:326-348.

Li AA, Baum MJ, McIntosh LJ, Day M, Liu F, Gray LE, Jr. 2008. Building a scientific framework for studying hormonal effects on behavior and on the development of the sexually dimorphic nervous system. Neurotoxicology 29:504-519.

OECD. 1998. Number 1. OECD Principles on Good Laboratory Practice (as revised in 1997). ENV/MC/CHEM(98)17. Available at http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/mc/chem(98)17&doclanguage=en

OECD. 2002a. Environment, Health and Safety Publications. Series on Testing and Assessment No. 35 and Series on Pesticides No. 14. Guidance notes for analysis and evaluation of chronic toxicity and carcinogenicity studies. ENV/JM/MONO(2002)19. Available at http://www.oecd.org/env/ehs/testing/seriesontestingandassessmenttestingforhumanhealth.htm

OECD. 2002b. OECD Series on Testing and Assessment Number 32 and OECD Series on Pesticides Number 10. Guidance Notes for Analysis and Evaluation of Repeat-Dose Toxicity Studies. ENV/JM/MONO(2000)18. Available at http://www.oecd.org/env/ehs/testing/seriesontestingandassessmenttestingforhumanhealth.htm

OECD. 2007a. Test No. 426: Developmental Neurotoxicity Study, Available at http://www.oecd-ilibrary.org/environment/oecd-guidelines-for-the-testing-of-chemicals-section-4-health-effects_20745788

OECD. 2007b. Test No. 440: Uterotrophic Bioassay in Rodents. Available at http://www.oecd-ilibrary.org/environment/oecd-guidelines-for-the-testing-of-chemicals-section-4-health-effects_20745788

OECD. 2008a. Test No. 407: Repeated Dose 28-day Oral Toxicity Study in Rodents, Available at http://www.oecd-ilibrary.org/environment/oecd-guidelines-for-the-testing-of-chemicals-section-4-health-effects_20745788

OECD. 2008b. Series on testing and assessment. Number 43. Guidance document on mammalian reproductive toxicity testing and assessment. ENV/JM/MONO(2008)16. Available at http://www.oecd.org/env/ehs/testing/seriesontestingandassessmenttestingforhumanhealth.htm

OECD. 2009a. Test No. 413: Subchronic Inhalation Toxicity: 90-day Study. Available at http://www.oecd-ilibrary.org/environment/oecd-guidelines-for-the-testing-of-chemicals-section-4-health-effects_20745788

OECD. 2009b. Test No. 451: Carcinogenicity Studies. Available at http://www.oecd-ilibrary.org/environment/oecd-guidelines-for-the-testing-of-chemicals-section-4-health-effects_20745788

OECD. 2009c. Test No. 452: Chronic Toxicity Studies. Available at http://www.oecd-ilibrary.org/environment/oecd-guidelines-for-the-testing-of-chemicals-section-4-health-effects_20745788

OECD. 2011. Test No. 443: Extended One-Generation Reproductive Toxicity Study. Available at http://www.oecd-ilibrary.org/environment/oecd-guidelines-for-the-testing-of-chemicals-section-4-health-effects_20745788

OECD. 2015. Test No. 421: Reproduction/Developmental Toxicity Screening Test. Available at http://www.oecd-ilibrary.org/environment/oecd-guidelines-for-the-testing-of-chemicals-section-4-health-effects_20745788