B.71 *IN VITRO* SKIN SENSITISATION ASSAYS ADDRESSING THE KEY EVENT ON ACTIVATION OF DENDRITIC CELLS ON THE ADVERSE OUTCOME PATHWAY (AOP) FOR SKIN SENSITISATION

GENERAL INTRODUCTION

Activation of dendritic cells key event based test method

1. A skin sensitiser refers to a substance that will lead to an allergic response following skin contact as defined by the United Nations Globally Harmonized System of Classification and Labelling of Chemicals (UN GHS) (1) and the European Union (EU) Regulation 1272/2008 on Classification, Labelling and Packaging of Substances and Mixtures (CLP)[1]. There is general agreement on the key biological events underlying skin sensitisation. The current knowledge of the chemical and biological mechanisms associated with skin sensitisation has been summarised as an Adverse Outcome Pathway (AOP) under the OECD AOP programme (2), starting with the molecular initiating event through intermediate events to the adverse effect, namely allergic contact dermatitis. In this instance, the molecular initiating event (i.e. the first key event) is the covalent binding of electrophilic substances to nucleophilic centres in skin proteins. The second key event in this AOP takes place in the keratinocytes and includes inflammatory responses as well as changes in gene expression associated with specific cell signalling pathways such as the antioxidant/electrophile response element (ARE)-dependent pathways. The third key event is the activation of dendritic cells (DC), typically assessed by expression of specific cell surface markers, chemokines and cytokines. The fourth key event is T-cell activation and proliferation, which is indirectly assessed in the murine Local Lymph Node Assay (LLNA) (3).

2. This test method (TM) is equivalent to OECD test guideline (TG) 442E (2017). It describes *in vitro* assays that address mechanisms described under the key event on activation of dendritic cells of the AOP for skin sensitisation (2). The TM comprises tests to be used for supporting the discrimination between skin sensitisers and non-sensitisers in accordance with the UN GHS and CLP.

[1] 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, OJ L 353/1, 31.12.2008

The tests described in this TM are:

- Human Cell Line Activation Test (h-CLAT)
- U937 cell line activation Test (U-SENS™)
- Interleukin-8 Reporter Gene Assay (IL-8 Luc assay)

3. The tests included in this test method and the corresponding OECD TG may differ in relation to the procedure used to generate the data and the readouts measured but can be used indiscriminately to address countries' requirements for test results on the Key Event on activation of dendritic cells of the AOP for skin sensitisation while benefiting from the OECD Mutual Acceptance of Data.

Background and principles of the tests included in the key event based test method

4. The assessment of skin sensitisation has typically involved the use of laboratory animals. The classical methods that use guinea-pigs, the Guinea Pig Maximisation Test (GPMT) of Magnusson and Kligman, and the Buehler Test (TM B.6) (4), assess both the induction and elicitation phases of skin sensitisation. The murine tests, the LLNA (TM B.42) (3) and its two non-radioactive modifications, LLNA: DA (TM B.50) (5) and LLNA: BrdU-ELISA (TM B.51) (6), all assess the induction response exclusively, and have also gained acceptance, since they provide an advantage over the guinea pig tests in terms of animal welfare together with an objective measurement of the induction phase of skin sensitisation.

5. Recently mechanistically-based *in chemico* and *in vitro* test methods addressing the first key event (TM B.59; Direct Peptide Reactivity Assay (7)), and second key event (TM B.60; ARE-Nrf2 Luciferase Test Method (8)) of the skin sensitisation AOP have been adopted for contributing to the evaluation of the skin sensitisation hazard potential of chemicals.

6. Tests described in this test method either quantify the change in the expression of cell surface marker(s) associated with the process of activation of monocytes and DC following exposure to sensitisers (e.g. CD54, CD86) or the changes in IL-8 expression, a cytokine associated with the activation of DC. Skin sensitisers have been reported to induce the expression of cell membrane markers such as CD40, CD54, CD80, CD83, and CD86 in addition to induction of proinflammatory cytokines, such as IL-1β and TNF-α, and several chemokines including IL-8 (CXCL8) and CCL3 (9) (10) (11) (12), associated with DC activation (2).

7. However, as DC activation represents only one key event of the skin sensitisation AOP (2) (13), information generated with tests measuring markers of DC activation alone may not be sufficient to conclude on the presence or absence of skin sensitisation potential of chemicals. Therefore data generated with the tests described in this test method are

proposed to support the discrimination between skin sensitisers (i.e. UN GHS/CLP Category 1) and non-sensitisers when used within Integrated Approaches to Testing and Assessment (IATA), together with other relevant complementary information, e.g. derived from *in vitro* assays addressing other key events of the skin sensitisation AOP as well as non-testing methods, including read-across from chemical analogues (13). Examples of the use of data generated with these tests within Defined Approaches, i.e. approaches standardised both in relation to the set of information sources used and in the procedure applied to the data to derive predictions, have been published (13) and can be employed as useful elements within IATA.

8. The tests described in this test method cannot be used on their own, neither to sub-categorise skin sensitisers into subcategories 1A and 1B as defined by UN GHS/CLP, for authorities implementing these two optional subcategories, nor to predict potency for safety assessment decisions. However, depending on the regulatory framework, positive results generated with these methods may be used on their own to classify a chemical into UN GHS/CLP category 1.

9. The term "test chemical" is used in this test method to refer to what is being tested[1] and is not related to the applicability of the tests to the testing of mono-constituent substances, multi-constituent substances and/or mixtures. Limited information is currently available on the applicability of the tests to multi-constituent substances/mixtures (14) (15). The tests are nevertheless technically applicable to the testing of multi-constituent substances and mixtures. However, before use of this test method on a mixture for generating data for an intended regulatory purpose, it should be considered whether, and if so why, it may provide adequate results for that purpose[2]. Such considerations are not needed when there is a regulatory requirement for the testing of the mixture. Moreover, when testing multi-constituent substances or mixtures, consideration should be given to possible interference of cytotoxic constituents with the observed responses.

[1] In June 2013, the OECD Joint Meeting agreed that where possible, a more consistent use of the term "test chemical" describing what is being tested should be applied in new and updated OECD test guidelines.

[2] This sentence was proposed and agreed at the April 2014 WNT meeting

LITERATURE

(1) United Nations UN (2015). Globally Harmonized System of Classification and Labelling of Chemicals (GHS). Sixth revised edition. New York & Geneva: United Nations Publications. ISBN: 978-92-1-117006-1. Available at: https://www.unece.org/trans/danger/publi/ghs/ghs_rev06/06files_e.html.

(2) OECD (2012). The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins. Part 1: Scientific Evidence. Series on Testing and Assessment No. 168. Available at: http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV/JM/MONO(2012)10/PART1&docLanguage=En.

(3) Chapter B.42 of this Annex: The Local Lymph Node Assay. Chapter B.6 of this Annex: Skin Sensitisation.

(4) Chapter B.50 of this Annex: Skin Sensitisation: Local Lymph Node Assay: DA.

(5) Chapter B.51 of this Annex: Skin sensitisation: Local Lymph Node Assay: BrdU-ELISA.

(6) Chapter B.59 of this Annex: In Chemico Skin Sensitisation: Direct Peptide Reactivity Assay (DPRA).

(7) Chapter B.60 of this Annex: *In Vitro* Skin Sensitisation: ARE-Nrf2 Luciferase Test Method.

(8) Steinman RM. (1991). The dendritic cell system and its role in immunogenicity. Annu Rev Immunol 9:271-96.

(9) Caux C, Vanbervliet B, Massacrier C, Azuma M, Okumura K, Lanier LL, and Banchereau J. (1994). B70/B7-2 is identical to CD86 and is the major functional ligand for CD28 expressed on human dendritic cells. J Exp Med 180:1841-7.

(10) Aiba S, Terunuma A, Manome H, and Tagami H. (1997). Dendritic cells differently respond to haptens and irritants by their production of cytokines and expression of co-stimulatory molecules. Eur J Immunol 27:3031-8.

(11) Aiba S, Manome H, Nakagawa S, Mollah ZU, Mizuashi M, Ohtani T, Yoshino Y, and Tagami. H. (2003). p38 mitogen-activated protein kinase and extracellular signal-regulated kinases play distinct roles in the activation of dendritic cells by two representative haptens, NiCl2 and DNCB. J Invest Dermatol 120:390-8.

(12) OECD (2016). Series on Testing & Assessment No 256: Guidance Document On The Reporting Of Defined Approaches And Individual

Information Sources To Be Used Within Integrated Approaches To Testing And Assessment (IATA) For Skin Sensitisation, Annex 1 and Annex 2. ENV/JM/HA(2016)29. Organisation for Economic Cooperation and Development, Paris. Available at: https://community.oecd.org/community/iatass.

(13) Ashikaga T, Sakaguchi H, Sono S, Kosaka N, Ishikawa M, Nukada Y, Miyazawa M, Ito Y, NishiyamaN, Itagaki H. (2010). A comparative evaluation of *in vitro* skin sensitisation tests: the human cell-line activation test (h-CLAT) versus the local lymph node assay (LLNA). Altern. Lab. Anim. 38, 275-284.

(14) Piroird, C., Ovigne, J.M., Rousset, F., Martinozzi-Teissier, S., Gomes, C., Cotovio, J., Alépée, N. (2015). The Myeloid U937 Skin Sensitization Test (U-SENS) addresses the activation of dendritic cell event in the adverse outcome pathway for skin sensitization. *Toxicol. In Vitro* 29, 901-916.

Appendix 1

IN VITRO SKIN SENSITISATION: HUMAN CELL LINE ACTIVATION TEST (H-CLAT)

INITIAL CONSIDERATIONS AND LIMITATIONS

1. The h-CLAT quantifies changes in the expression of cell surface markers associated with the process of activation of monocytes and dendritic cells (DC) (i.e. CD86 and CD54), in the human monocytic leukaemia cell line THP-1, following exposure to sensitisers (1)(2). The measured expression levels of CD86 and CD54 cell surface markers are then used for supporting the discrimination between skin sensitisers and non-sensitisers.

2. The h-CLAT has been evaluated in a European Union Reference Laboratory for Alternatives to Animal Testing (EURL ECVAM)-coordinated validation study and subsequent independent peer review by the EURL ECVAM Scientific Advisory Committee (ESAC). Considering all available evidence and input from regulators and stakeholders, the h-CLAT was recommended by EURL ECVAM (3) to be used as part of an IATA to support the discrimination between sensitisers and non-sensitisers for the purpose of hazard classification and labelling. Examples of the use of h-CLAT data in combination with other information are reported in the literature (4)(5)(6)(7)(8)(9)(10)(11).

3. The h-CLAT proved to be transferable to laboratories experienced in cell culture techniques and flow cytometry analysis. The level of reproducibility in predictions that can be expected from the test is in the order of 80% within and between laboratories (3)(12). Results generated in the validation study (13) and other published studies (14) overall indicate that, compared with LLNA results, the accuracy in distinguishing skin sensitisers (i.e. UN GHS/CLP Cat.1) from non-sensitisers is 85% (N=142) with a sensitivity of 93% (94/101) and a specificity of 66% (27/41) (based on a re-analysis by EURL ECVAM (12) considering all existing data and not considering negative results for chemicals with a Log Kow greater than 3.5 as described in paragraph 4). False negative predictions with the h-CLAT are more likely to concern chemicals showing a low to moderate skin sensitisation potency (i.e. UN GHS/CLP subcategory 1B) than chemicals showing a high skin sensitisation potency (i.e. UN GHS/CLP subcategory 1A) (4)(13)(15). Taken together, this information indicates the usefulness of the h-CLAT method to contribute to the identification of skin sensitisation hazards. However, the accuracy values given here for the h-CLAT as a stand-alone test are only indicative, since the test should be considered in combination with other sources of information in the context of an IATA and in accordance with the provisions of paragraphs 7 and 8 in the General introduction. Furthermore, when evaluating non-animal methods for skin sensitisation, it should be kept in mind that the LLNA test as well as other animal tests may not fully reflect the situation in humans.

4. On the basis of the data currently available, the h-CLAT method was shown to be applicable to test chemicals covering a variety of organic functional groups, reaction mechanisms, skin sensitisation potency (as determined in *in vivo* studies) and physicochemical properties (3)(14)(15). The h-CLAT method is applicable to test chemicals soluble or that form a stable dispersion (i.e. a colloid or suspension in which the test chemical does not settle or separate from the solvent/vehicle into different phases) in an appropriate solvent/vehicle (see paragraph 14). Test chemicals with a Log Kow greater than 3.5 tend to produce false negative results (14). Therefore negative results with test chemicals with a Log Kow greater than 3.5 should not be considered. However, positive results obtained with test chemicals with a Log Kow greater than 3.5 could still be used to support the identification of the test chemical as a skin sensitiser. Furthermore, because of the limited metabolic capability of the cell line used (16) and because of the experimental conditions, pro-haptens (i.e. substances requiring enzymatic activation for example via P450 enzymes) and pre-haptens (i.e. substances activated by oxidation) in particular with a slow oxidation rate may also provide negative results in the h-CLAT (15). Fluorescent test chemicals can be assessed with the h-CLAT (17), nevertheless, strong fluorescent test chemicals emitting at the same wavelength as fluorescein isothiocyanate (FITC) or as propidium iodide (PI), will interfere with the flow cytometric detection and thus cannot be correctly evaluated using FITC-conjugated antibodies or PI. In such a case, other fluorochrome-tagged antibodies or other cytotoxicity markers, respectively, can be used as long as it can be shown they provide similar results as the FITC-tagged antibodies (see paragraph 24) or PI (see paragraph 18) e.g. by testing the proficiency substances in Appendix 1-2. In the light of the above, negative results should be interpreted in the context of the stated limitations and together with other information sources within the framework of IATA. In cases where there is evidence demonstrating the non-applicability of the h-CLAT method to other specific categories of test chemicals, it should not be used for those specific categories.

5. As described above, the h-CLAT method supports the discrimination between skin sensitisers from non-sensitisers. However, it may also potentially contribute to the assessment of sensitising potency (4)(5)(9) when used in integrated approaches such as IATA. Nevertheless, further work, preferably based on human data, is required to determine how h-CLAT results may possibly inform potency assessment.

6. Definitions are provided in Appendix 1.1.

PRINCIPLE OF THE TEST

7. The h-CLAT method is an *in vitro* assay that quantifies changes of cell surface marker expression (i.e. CD86 and CD54) on a human monocytic leukemia cell line, THP-1 cells, following 24 hours exposure to the test chemical. These surface molecules are typical markers of monocytic THP-1 activation and may mimic DC activation, which plays a

critical role in T-cell priming. The changes of surface marker expression are measured by flow cytometry following cell staining with fluorochrome-tagged antibodies. Cytotoxicity measurement is also conducted concurrently to assess whether upregulation of surface marker expression occurs at sub-cytotoxic concentrations. The relative fluorescence intensity of surface markers compared to solvent/vehicle control are calculated and used in the prediction model (see paragraph 26), to support the discrimination between sensitisers and non-sensitisers

DEMONSTRATION OF PROFICIENCY

8. Prior to routine use of the test described in this Appendix to test method B.71, laboratories should demonstrate technical proficiency, using the 10 proficiency substances listed in Appendix 1.2. Moreover, test users should maintain an historical database of data generated with the reactivity checks (see paragraph 11) and with the positive and solvent/vehicle controls (see paragraphs 20-22), and use these data to confirm the reproducibility of the test in their laboratory is maintained over time.

PROCEDURE

9. This test is based on the h-CLAT DataBase service on ALternative Methods to animal experimentation (DB-ALM) protocol no 158 (18) which represents the protocol used for the EURL ECVAM-coordinated validation study. It is recommended that this protocol is used when implementing and using the h-CLAT method in the laboratory. The following is a description of the main components and procedures for the h-CLAT method, which comprises two steps: *dose finding assay* and *CD86/CD54 expression measurement*.

Preparation of cells

10. The human monocytic leukaemia cell line, THP-1, should be used for performing the h-CLAT method. It is recommended that cells (TIB-202™) are obtained from a well-qualified cell bank, such as the American Type Culture Collection.

11. THP-1 cells are cultured, at 37°C under 5% CO_2 and humidified atmosphere, in RPMI-1640 medium supplemented with 10% foetal bovine serum (FBS), 0.05 mM 2-mercaptoethanol, 100 units/ml penicillin and 100 µg/ml streptomycin. The use of penicillin and streptomycin in the culture medium can be avoided. However, in such a case users should verify that the absence of antibiotics in the culture medium has no impact on the results, for example by testing the proficiency substances listed in Appendix 1.2. In any case, in order to minimise the risk of contamination, good cell culture practices should be followed independently of the presence or not of antibiotics in the cell culture medium. THP-1 cells are routinely seeded every 2-3 days at the density of 0.1 to 0.2 × 10^6 cells/ml. They should be maintained at densities from 0.1 to 1.0 × 10^6 cells/ml. Prior to using them for testing, the cells should be qualified by conducting a reactivity check. The reactivity

check of the cells should be performed using the positive controls, 2,4-dinitrochlorobenzene (DNCB) (CAS no 97-00-7, ≥ 99% purity) and nickel sulfate ($NiSO_4$) (CAS no 10101-97-0, ≥ 99% purity) and the negative control, lactic acid (LA) (CAS no 50-21-5, ≥ 85% purity), two weeks after thawing. Both DNCB and $NiSO_4$ should produce a positive response of both CD86 and CD54 cell surface markers, and LA should produce a negative response of both CD86 and CD54 cell surface markers. Only the cells which passed the reactivity check are to be used for the assay. Cells can be propagated up to two months after thawing. Passage number should not exceed 30. The reactivity check should be performed according to the procedures described in paragraphs 20-24.

12. For testing, THP-1 cells are seeded at a density of either 0.1×10^6 cells/ml or 0.2×10^6 cells/ml, and pre-cultured in culture flasks for 72 hours or for 48 hours, respectively. It is important that the cell density in the culture flask just after the pre-culture period be as consistent as possible in each experiment (by using one of the two pre-culture conditions described above), because the cell density in the culture flask just after pre-culture could affect the CD86/CD54 expression induced by allergens (19). On the day of testing, cells harvested from culture flask are resuspended with fresh culture medium at 2×10^6 cells/ml. Then, cells are distributed into a 24 well flat-bottom plate with 500 µl (1×10^6 cells/well) or a 96-well flat-bottom plate with 80 µl (1.6×10^5 cells/well).

Dose finding assay

13. A *dose finding assay* is performed to determine the CV75, being the test chemical concentration that results in 75% cell viability (CV) compared to the solvent/vehicle control. The CV75 value is used to determine the concentration of test chemicals for the *CD86/CD54* expression measurement (see paragraphs 20-24).

Preparation of test chemicals and control substances

14. The test chemicals and control substances are prepared on the day of testing. For the h-CLAT method, test chemicals are dissolved or stably dispersed (see also paragraph 4) in saline or medium as first solvent/vehicle options or dimethyl sulfoxide (DMSO, ≥ 99% purity) as a second solvent/vehicle option if the test chemical is not soluble or does not form a stable dispersion in the previous two solvents/vehicles, to final concentrations of 100 mg/ml (in saline or medium) or 500 mg/ml (in DMSO). Other solvents/vehicles than those described above may be used if sufficient scientific rationale is provided. Stability of the test chemical in the final solvent/vehicle should be taken into account.

15. Starting from the 100 mg/ml (in saline or medium) or 500 mg/ml (in DMSO) stock solutions of the test chemicals, the following dilution steps should be taken:

- For saline or medium as solvent/vehicle: Eight stock solutions (eight concentrations) are prepared, by two-fold serial dilutions using the corresponding solvent/vehicle. These stock solutions are then further diluted 50-fold into culture medium (working solutions). If the

top final concentration in the plate of 1000 μg/ml is non-toxic, the maximum concentration should be re-determined by performing a new cytotoxicity test. The final concentration in the plate should not exceed 5000 μg/ml for test chemicals dissolved or stably dispersed in saline or medium.

- For DMSO as solvent/vehicle: Eight stock solutions (eight concentrations) are prepared, by two-fold serial dilutions using the corresponding solvent/vehicle. These stock solutions are then further diluted 250-fold into culture medium (working solutions).The final concentration in plate should not exceed 1000 μg/ml even if this concentration is non-toxic.

The working solutions are finally used for exposure by adding an equal volume of working solution to the volume of THP-1 cell suspension in the plate (see also paragraph 17) to achieve a further two-fold dilution (usually, the final range of concentrations in the plate is 7.81–1000 μg/ml).

16. The solvent/vehicle control used in the h-CLAT method is culture medium (for test chemicals solubilised or stably dispersed (see paragraph 4) either with medium or saline) or DMSO (for test chemicals solubilised or stably dispersed in DMSO) tested at a single final concentration in the plate of 0.2%. It undergoes the same dilution as described for the working solutions in paragraph 15.

Application of test chemicals and control substances

17. The culture medium or working solutions described in paragraphs 15 and 16 are mixed 1:1 (v/v) with the cell suspensions prepared in the 24-well or 96-well flat-bottom plate (see paragraph 12). The treated plates are then incubated for 24±0.5 hours at 37°C under 5% CO_2. Care should be taken to avoid evaporation of volatile test chemicals and cross-contamination between wells by test chemicals, e.g. by sealing the plate prior to the incubation with the test chemicals (20).

Propidium iodide (PI) staining

18. After 24±0.5 hours of exposure, cells are transferred into sample tubes and collected by centrifugation. The supernatants are discarded and the remaining cells are resuspended with 200 μl (in case of 96-well) or 600 μl (in case of 24-well) of a phosphate buffered saline containing 0.1% bovine serum albumin (staining buffer). 200 μl of cell suspension is transferred into 96-well round-bottom plate (in case of 96-well) or micro tube (in case of 24-well) and washed twice with 200 μl (in case of 96-well) or 600 μl (in case of 24-well) of staining buffer. Finally, cells are resuspended in staining buffer (e.g. 400 μl) and PI solution (e.g. 20 μl) is added (for example, final concentration of PI is 0.625 μg/ml). Other cytotoxicity markers, such as 7-Aminoactinomycin D (7-AAD), Trypan blue or others may be used if the alternative stains can be shown to provide similar results as PI, for example by testing the proficiency substances in Appendix 1.2.

Cytotoxicity measurement by flow cytometry and estimation of CV75 value

19. The PI uptake is analysed using flow cytometry with the acquisition channel FL-3. A total of 10 000 living cells (PI negative) are acquired. The cell viability can be calculated using the following equation by the cytometer analysis program. When the cell viability is low, up to 30 000 cells including dead cells should be acquired. Alternatively, data can be acquired for one minute after the initiation of the analysis.

$$Cell\ viability = \frac{Number\ of\ living\ cells}{Total\ Number\ of\ aquired\ cells} \times 100$$

The CV75 value (see paragraph 13), i.e. a concentration showing 75% of THP-1 cell survival (25% cytotoxicity), is calculated by log-linear interpolation using the following equation:

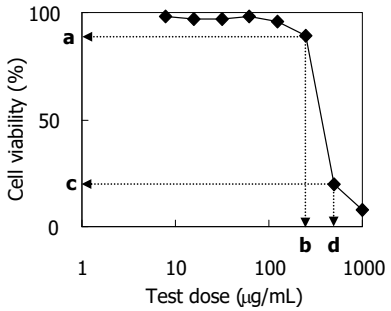
$$Log\ CV75 = \frac{(75-c)\times Log(b)-(75-a)\times Log(d)}{a-c}$$

Where:

a is the minimum value of cell viability over 75%

c is the maximum value of cell viability below 75%

b and d are the concentrations showing the value of cell viability a and c respectively



Other approaches to derive the CV75 can be used as long as it is demonstrated that this has no impact on the results (e.g. by testing the proficiency substances).

CD86/CD54 expression measurement

Preparation of the test chemicals and control substances

20. The appropriate solvent/vehicle (saline, medium or DMSO; see paragraph 14) is used to dissolve or stably disperse the test chemicals. The test chemicals are first diluted to the concentration corresponding to 100-fold (for saline or medium) or 500-fold (for DMSO) of the 1.2 × CV75 determined in the *dose finding assay* (see paragraph 19). If the CV75 cannot be determined (i.e. if sufficient cytotoxicity is not observed in the *dose finding assay*), the highest soluble or stably dispersed concentration of test chemical prepared with each solvent/vehicle should be used as starting concentration. Please note that the final concentration in the plate should not exceed 5000 μg/ml (in case of saline or medium) or

1000 µg/ml (in case of DMSO). Then, 1.2-fold serial dilutions are made using the corresponding solvent/vehicle to obtain the stock solutions (eight concentrations ranging from 100×1.2 × CV75 to 100×0.335 × CV75 (for saline or medium) or from 500×1.2 × CV75 to 500×0.335 × CV75 (for DMSO)) to be tested in the h-CLAT method (see DB-ALM protocol N_O. 158 for an example of dosing scheme). The stock solutions are then further diluted 50-fold (for saline or medium) or 250-fold (for DMSO) into the culture medium (working solutions). These working solutions are finally used for exposure with a further final two-fold dilution factor in the plate. If the results do not meet the acceptance criteria described in the paragraphs 29 and 30 regarding cell viability, the *dose finding assay* may be repeated to determine a more precise CV75. Please note that only 24-well plates can be used for CD86/CD54 expression measurement.

21. The solvent/vehicle control is prepared as described in paragraph 16. The positive control used in the h-CLAT method is DNCB (see paragraph 11), for which stock solutions are prepared in DMSO and diluted as described for the stock solutions in paragraph 20. DNCB should be used as the positive control for *CD86/CD54 expression measurement* at a final single concentration in the plate (typically 4.0 µg/ml). To obtain a 4.0 µg/ml concentration of DNCB in the plate, a 2 mg/ml stock solution of DNCB in DMSO is prepared and further diluted 250-fold with culture medium to a 8 µg/ml working solution. Alternatively, the CV75 of DNCB, which is determined in each test facility, could be also used as the positive control concentration. Other suitable positive controls may be used if historical data are available to derive comparable run acceptance criteria. For positive controls, the final single concentration in the plate should not exceed 5000 µg/ml (in case of saline or medium) or 1000 µg/ml (in case of DMSO). The run acceptance criteria are the same as those described for the test chemical (see paragraph 29), except for the last acceptance criterion since the positive control is tested at a single concentration.

Application of test chemicals and control substances

22. For each test chemical and control substance, one experiment is needed to obtain a prediction. Each experiment consists of at least two independent runs for *CD86/CD54 expression measurement* (see paragraphs 26-28). Each independent run is performed on a different day or on the same day provided that for each run: a) independent fresh stock solutions and working solutions of the test chemical and antibody solutions are prepared and b) independently harvested cells are used (i.e. cells are collected from different culture flasks); however, cells may come from the same passage. Test chemicals and control substances prepared as working solutions (500 µl) are mixed with 500 µl of suspended cells ($1x10^6$ cells) at 1:1 ratio, and cells are incubated for 24±0.5 hours as described in paragraphs 20 and 21. In each run, a single replicate for each concentration of the test chemical and control substance is sufficient because a prediction is obtained from at least two independent runs.

Cell staining and analysis

23. After 24±0.5 hours of exposure, cells are transferred from 24 well plate into sample tubes, collected by centrifugation and then washed twice with 1ml of staining buffer (if necessary, additional washing steps may be done). After washing, cells are blocked with 600 µl of blocking solution (staining buffer containing 0.01% (w/v) globulin (Cohn fraction II, III, human; SIGMA, #G2388-10G or equivalent)) and incubated at 4°C for 15 min. After blocking, cells are split in three aliquots of 180 µl into a 96-well round-bottom plate or micro tube.

24. After centrifugation, cells are stained with 50 µl of FITC-labelled anti-CD86, anti-CD54 or mouse IgG1 (isotype) antibodies at 4°C for 30 min. The antibodies described in the h-CLAT DB-ALM protocol no 158 (18) should be used by diluting 3:25 v/v (for CD86 (BD-PharMingen, #555657; Clone: Fun-1)) or 3:50 v/v (for CD54 (DAKO, #F7143; Clone: 6.5B5) and IgG1 (DAKO, #X0927)) with staining buffer. These antibody dilution factors were defined by the test developers as those providing the best signal-to-noise ratio. Based on the experience of the test developers, the fluorescence intensity of the antibodies is usually consistent between different lots. However, users may consider titrating the antibodies in their own laboratory's conditions to define the best concentrations for use. Other fluorochrome-tagged anti-CD86 and/or anti-CD54 antibodies may be used if they can be shown to provide similar results as FITC-conjugated antibodies, for example by testing the proficiency substances in Appendix 1.2. It should be noted that changing the clone or supplier of the antibodies as described in the h-CLAT DB-ALM protocol no 158 (18) may affect the results. After washing twice or more with 150 µl of staining buffer, cells are resuspended in staining buffer (e.g. 400 µl), and the PI solution (e.g. 20 µl to obtain a final concentration of 0.625 µg/ml) or another cytotoxicity marker's solution (see paragraph 18) is added. The expression levels of CD86 and CD54, and cell viability are analysed using flow cytometry.

DATA AND REPORTING

Data evaluation

25. The expression of CD86 and CD54 is analysed with flow cytometry with the acquisition channel FL-1. Based on the geometric mean fluorescence intensity (MFI), the relative fluorescence intensity (RFI) of CD86 and CD54 for positive control (ctrl) cells and chemical-treated cells are calculated according to the following equation:

$$RFI = \frac{MFI\ of\ chemical\text{-}treated\ cell - MFI\ of\ chemical\text{-}treated\ isotype\ control\ cells}{MFI\ of solvent/vehicle\text{-}treated\ ctrl\ cells - MFI\ of\ solvent/vehicle\text{-}treated\ isotype\ ctrl\ cells} \times 100$$

The cell viability from the isotype control (ctrl) cells (which are stained with mouse IgG1 (isotype) antibodies) is also calculated according to the equation described in paragraph 19.

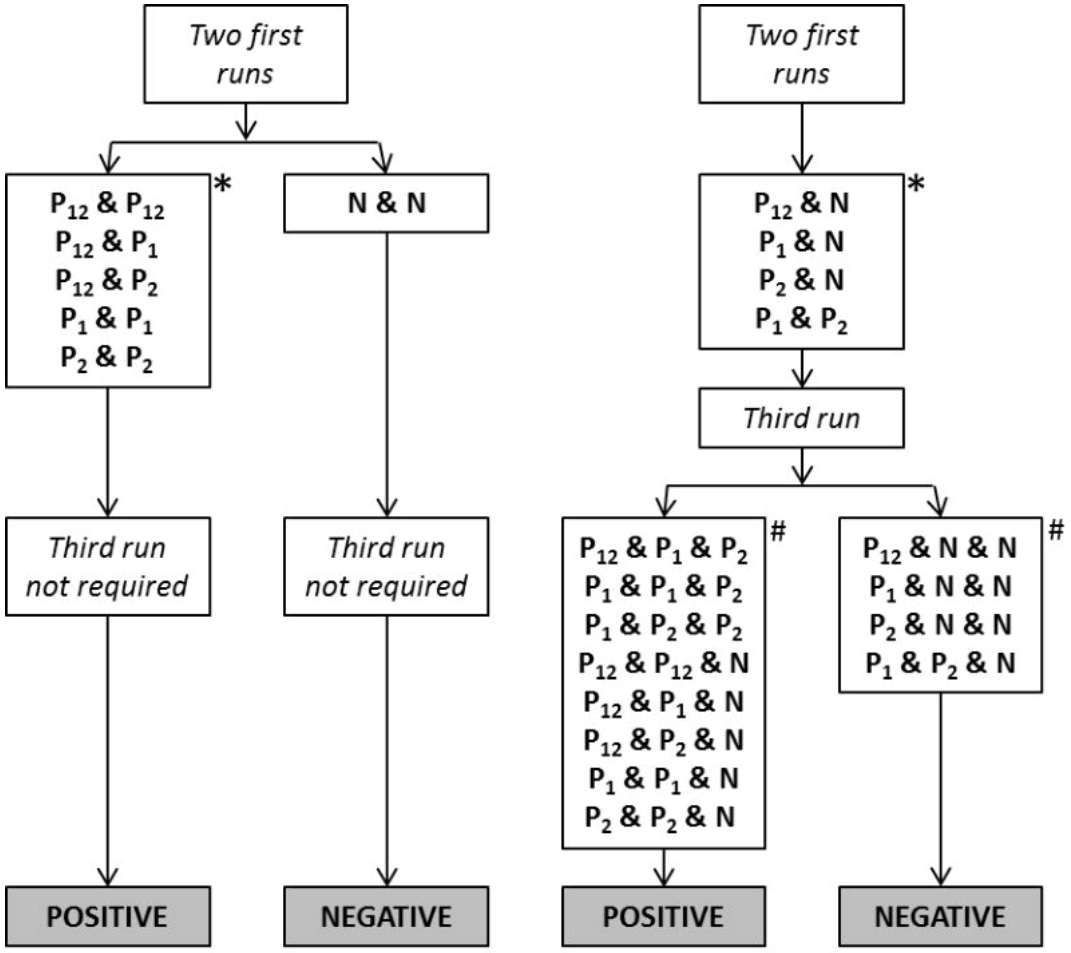
Prediction model

26. For *CD86/CD54* expression measurement, each test chemical is tested in at least two independent runs to derive a single prediction (POSITIVE or NEGATIVE). An h-CLAT prediction is considered POSITIVE if at least one of the following conditions is met in 2 of 2 or in at least 2 of 3 independent runs, otherwise the h-CLAT prediction is considered NEGATIVE (Figure 1):

- The RFI of CD86 is equal to or greater than 150% at any tested concentration (with cell viability ≥ 50%);
- The RFI of CD54 is equal to or greater than 200% at any tested concentration (with cell viability ≥ 50%).

27. Based on the above, if the first two runs are both positive for CD86 and/or are both positive for CD54, the h-CLAT prediction is considered POSITIVE and a third run does not need to be conducted. Similarly, if the first two runs are negative for both markers, the h-CLAT prediction is considered NEGATIVE (with due consideration of the provisions of paragraph 30) without the need for a third run. If however, the first two runs are not concordant for at least one of the markers (CD54 or CD86), a third run is needed and the final prediction will be based on the majority result of the three individual runs (i.e. 2 out of 3). In this respect, it should be noted that if two independent runs are conducted and one is only positive for CD86 (hereinafter referred to as P_1) and the other is only positive for CD54 (hereinafter referred to as P_2), a third run is required. If this third run is negative for both markers (hereinafter referred to as N), the h-CLAT prediction is considered NEGATIVE. On the other hand, if the third run is positive for either marker (P_1 or P_2) or for both markers (hereinafter referred to as P_{12}), the h-CLAT prediction is considered POSITIVE.

Figure 1: Prediction model used in the h-CLAT method. An h-CLAT prediction should be considered in the framework of an IATA and in accordance with the provision of paragraphs 7 and 8 in the General introduction.



P_1: run with only CD86 positive; P_2; run with only CD54 positive; P_{12}: run with both CD86 and CD54 positive; N: run with neither CD86 nor CD54 positive.

*The boxes show the relevant combinations of results from the first two runs, independently of the order in which they may be obtained.

#The boxes show the relevant combinations of results from the three runs on the basis of the results obtained in the first two runs shown in the box above, but do not reflect the order in which they may be obtained.

28. For the test chemicals predicted as POSITIVE with the h-CLAT, optionally, two Effective Concentrations (EC) values, the EC150 for CD86 and EC200 for CD54, i.e. the concentration at which the test chemicals induced a RFI of 150 or 200, may be determined. These EC values potentially could contribute to the assessment of sensitising potency (9) when used in integrated approaches such as IATA (4) (5) (6) (7) (8). They can be calculated by the following equations:

$$EC150\ (for\ CD86) = B_{conc} + [(150 - B_{RFI)}/(A_{RFI} - B_{RFI}) \times (A_{conc} - B_{conc})]$$

$$EC200\ (for\ CD54) = B_{conc} + [(200 - B_{RFI)}/(A_{RFI} - B_{RFI}) \times (A_{conc} - B_{conc})]$$

where

A_{conc} is the lowest concentration in μg/ml with RFI > 150 (CD86) or 200 (CD54)

B_{conc} is the highest concentration in μg/ml with RFI < 150 (CD86) or 200 (CD54)

A_{RFI} is the RFI at the lowest concentration with RFI > 150 (CD86) or 200 (CD54)

B_{RFI} is the RFI at the highest concentration with RFI < 150 (CD86) or 200 (CD54)

For the purpose of more precisely deriving the EC150 and EC200 values, three independent runs for *CD86/CD54 expression measurement* may be required. The final EC150 and EC200 values are then determined as the median value of the ECs calculated from the three independent runs. When only two of three independent runs meet the criteria for positivity (see paragraphs 26-27), the higher EC150 or EC200 of the two calculated values is adopted.

Acceptance criteria

29. The following acceptance criteria should be met when using the h-CLAT method (22) (27).

- The cell viabilities of medium and solvent/vehicle controls should be higher than 90%.
- In the solvent/vehicle control, RFI values of both CD86 and CD54 should not exceed the positive criteria (CD86 RFI ≥ 150% and CD54 RFI ≥ 200%). RFI values of the solvent/vehicle control are calculated by using the formula described in paragraph 25 ("MFI of chemical" should be replaced with "MFI of solvent/vehicle", and "MFI of solvent/vehicle" should be replaced with "MFI of (medium) control").
- For both medium and solvent/vehicle controls, the MFI ratio of both CD86 and CD54 to isotype control should be > 105%.
- In the positive control (DNCB), RFI values of both CD86 and CD54 should meet the positive criteria (CD86 RFI ≥ 150 and CD54 RFI ≥ 200) and cell viability should be more than 50%.
- For the test chemical, the cell viability should be more than 50% in at least four tested concentrations in each run.

30. Negative results are acceptable only for test chemicals exhibiting a cell viability of less than 90% at the highest concentration tested (i.e. 1.2 × CV75 according to the serial dilution scheme described in paragraph 20). If the cell viability at 1.2 × CV75 is equal or above 90% the negative result should be discarded. In such a case it is recommended to try to refine the dose selection by repeating the CV75 determination. It should be noted that when 5000 μg/ml in saline (or medium or other solvents/vehicles), 1000 μg/ml in DMSO or the highest soluble concentration is used as the maximal test concentration of a test chemical, a negative result is acceptable even if the cell viability is above 90%.

Test report

31. The test report should include the following information.

Test chemical

Mono-constituent substance

- Chemical identification, such as IUPAC or CAS name(s), CAS number(s), SMILES or InChI code, structural formula, and/or other identifiers;
- Physical appearance, Log Kow, water solubility, DMSO solubility, molecular weight, and additional relevant physicochemical properties, to the extent available;
- Purity, chemical identity of impurities as appropriate and practically feasible, etc.;
- Treatment prior to testing, if applicable (e.g. warming, grinding);
- Concentration(s) tested;
- Storage conditions and stability to the extent available;
- Justification for choice of solvent/vehicle for each test chemical.

Multi-constituent substance, UVCB and mixture

- Characterisation as far as possible by e.g. chemical identity (see above), purity, quantitative occurrence and relevant physicochemical properties (see above) of the constituents, to the extent available;
- Physical appearance, water solubility, DMSO solubility and additional relevant physicochemical properties, to the extent available;
- Molecular weight or apparent molecular weight in case of mixtures/polymers of known compositions or other information relevant for the conduct of the study;
- Treatment prior to testing, if applicable (e.g. warming, grinding);
- Concentration(s) tested;
- Storage conditions and stability to the extent available;
- Justification for choice of solvent/vehicle for each test chemical.

Controls

Positive control

- Chemical identification, such as IUPAC or CAS name(s), CAS number(s), SMILES or InChI code, structural formula, and/or other identifiers;
- Physical appearance, Log K_{ow}, water solubility, DMSO solubility, molecular weight, and additional relevant physicochemical properties, to the extent available and where applicable;
- Purity, chemical identity of impurities as appropriate and practically feasible, etc.;
- Treatment prior to testing, if applicable (e.g. warming, grinding);

- Concentration(s) tested;
- Storage conditions and stability to the extent available;
- Reference to historical positive control results demonstrating suitable run acceptance criteria, if applicable.

Negative and solvent/vehicle control

- Chemical identification, such as IUPAC or CAS name(s), CAS number(s), SMILES or InChI code, structural formula, and/or other identifiers;
- Purity, chemical identity of impurities as appropriate and practically feasible, etc.;
- Physical appearance, molecular weight, and additional relevant physicochemical properties in the case other control solvent/vehicle than those mentioned in the Test Guideline are used and to the extent available;
- Storage conditions and stability to the extent available;
- Justification for choice of solvent/vehicle for each test chemical.

Test conditions

- Name and address of the sponsor, test facility and study director;
- Description of testused;
- Cell line used, its storage conditions and source (e.g. the facility from which they were obtained);
- Flow cytometry used (e.g. model), including instrument settings, globulin, antibodies and cytotoxicity marker used;
- The procedure used to demonstrate proficiency of the laboratory in performing the test by testing of proficiency substances, and the procedure used to demonstrate reproducible performance of the test over time, e.g. historical control data and/or historical reactivity checks' data.

Test acceptance criteria

- Cell viability, MFI and RFI values obtained with the solvent/vehicle control in comparison to the acceptance ranges;
- Cell viability and RFI values obtained with the positive control in comparison to the acceptance ranges;
- Cell viability of all tested concentrations of the tested chemical.

Test procedure

- Number of runs used;

- Test chemical concentrations, application and exposure time used (if different than the one recommended)
- Duration of exposure (if different than the one recommended);
- Description of evaluation and decision criteria used;
- Description of any modifications of the test procedure.

Results

- Tabulation of the data, including CV75 (if applicable), individual geometric MFI, RFI, cell viability values, EC150/EC200 values (if applicable) obtained for the test chemical and for the positive control in each run, and an indication of the rating of the test chemical according to the prediction model;
- Description of any other relevant observations, if applicable.

Discussion of the results

- Discussion of the results obtained with the h-CLAT method;
- Consideration of the test results within the context of an IATA, if other relevant information is available.

Conclusions

LITERATURE

(1) Ashikaga T, Yoshida Y, Hirota M, Yoneyama K, Itagaki H, Sakaguchi H, Miyazawa M, Ito Y, Suzuki H, Toyoda H. (2006). Development of an *in vitro* skin sensitization test using human cell lines: The human Cell Line Activation Test (h-CLAT) I. Optimization of the h-CLAT protocol. Toxicol. *In Vitro* 20, 767–773.

(2) Miyazawa M, Ito Y, Yoshida Y, Sakaguchi H, Suzuki H. (2007). Phenotypic alterations and cytokine production in THP-1 cells in response to allergens. Toxicol. *In Vitro* 21, 428-437.

(3) EC EURL-ECVAM (2013). Recommendation on the human Cell Line Activation Test (h-CLAT) for skin sensitisation testing. Accessible at: https://eurl-ecvam.jrc.ec.europa.eu/eurl-ecvam-recommendations

(4) Takenouchi O, Fukui S, Okamoto K, Kurotani S, Imai N, Fujishiro M, Kyotani D, Kato Y, Kasahara T, Fujita M, Toyoda A, Sekiya D, Watanabe S, Seto H, Hirota M, Ashikaga T, Miyazawa M. (2015). Test battery with the human cell line activation test, direct peptide reactivity assay and DEREK based on a 139 chemical data set for predicting skin sensitizing potential and potency of chemicals. J Appl Toxicol. 35, 1318-1332.

(5) Hirota M, Fukui S, Okamoto K, Kurotani S, Imai N, Fujishiro M, Kyotani D, Kato Y, Kasahara T, Fujita M, Toyoda A, Sekiya D, Watanabe S, Seto H, Takenouchi O, Ashikaga T, Miyazawa M. (2015). Evaluation of combinations of *in vitro* sensitization test descriptors for the artificial neural network-based risk assessment model of skin sensitization. J Appl Toxicol. 35, 1333-1347.

(6) Bauch C, Kolle SN, Ramirez T, Fabian E, Mehling A, Teubner W, van Ravenzwaay B, Landsiedel R. (2012). Putting the parts together: combining *in vitro* methods to test for skin sensitizing potencials. Regul Toxicol Parmacol. 63, 489-504.

(7) Van der Veen JW, Rorije E, Emter R, Natch A, van Loveren H, Ezendam J. (2014). Evaluating the performance of integrated approaches for hazard identification of skin sensitizing chemicals. Regul Toxicol Pharmacol. 69, 371-379.

(8) Urbisch D, Mehling A, Guth K, Ramirez T, Honarvar N, Kolle S, Landsiedel R, Jaworska J, Kern PS, Gerberick F, Natsch A, Emter R, Ashikaga T, Miyazawa M, Sakaguchi H. (2015). Assessing skin sensitization hazard in mice and men using non-animal test methods. Regul Toxicol Parmacol. 71, 337-351.

(9) Jaworska JS, Natsch A, Ryan C, Strickland J, Ashikaga T, Miyazawa M. (2015). Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment: a decision support system for quantitative weight of evidence and adaptive testing strategy. Arch Toxicol. 89, 2355-2383.

(10) Strickland J, Zang Q, Kleinstreuer N, Paris M, Lehmann DM, Choksi N, Matheson J, Jacobs A, Lowit A, Allen D, Casey W. (2016). Integrated decision strategies for skin sensitization hazard. J Appl Toxicol. DOI 10.1002/jat.3281.

(11) Nukada Y, Ashikaga T, Miyazawa M, Hirota M, Sakaguchi H, Sasa H, Nishiyama N. (2012). Prediction of skin sensitization potency of chemicals by human Cell Line Activation Test (h-CLAT) and an attempt at classifying skin sensitization potency. Toxicol. *In Vitro* 26, 1150-60.

(12) EC EURL ECVAM (2015). Re-analysis of the within and between laboratory reproducibility of the human Cell Line Activation Test (h-CLAT). Accessible at: https://eurl-ecvam.jrc.ec.europa.eu/eurl-ecvam-recommendations/eurl-ecvam-recommendation-on-the-human-cell-line-activation-test-h-clat-for-skin-sensitisation-testing

(13) EC EURL ECVAM (2012). human Cell Line Activation Test (h-CLAT) Validation Study Report Accessible at: https://eurl-ecvam.jrc.ec.europa.eu/eurl-ecvam-recommendations

(14) Takenouchi O, Miyazawa M, Saito K, Ashikaga T, Sakaguchi H. (2013). Predictive performance of the human Cell Line Activation Test (h-CLAT) for lipophilic with high octanol-water partition coefficients. J. Toxicol. Sci. 38, 599-609.

(15) Ashikaga T, Sakaguchi H, Sono S, Kosaka N, Ishikawa M, Nukada Y, Miyazawa M, Ito Y, NishiyamaN, Itagaki H. (2010). A comparative evaluation of *in vitro* skin sensitisation tests: the human cell-line activation test (h-CLAT) versus the local lymph node assay (LLNA). Altern. Lab. Anim. 38, 275-284.

(16) Fabian E., Vogel D., Blatz V., Ramirez T., Kolle S., Eltze T., van Ravenzwaay B., Oesch F., Landsiedel R. (2013). Xenobiotic metabolizin enzyme activities in cells used for testing skin sensitization *in vitro*. Arch Toxicol 87, 1683-1969.

(17) Okamoto K, Kato Y, Kosaka N, Mizuno M, Inaba H, Sono S, Ashikaga T, Nakamura T, Okamoto Y, Sakaguchi H, Kishi M, Kuwahara H, Ohno Y. (2010). The Japanese ring study of a human Cell Line Activation Test (h-CLAT) for predicting skin sensitization potential (6^{th} report): A study for evaluating oxidative hair dye sensitization potential using h-CLAT. AATEX 15, 81-88.

(18) DB-ALM (INVITTOX) (2014). Protocol 158: human Cell Line Activation Test (h-CLAT), 23pp. Accessible at: http://ecvam-dbalm.jrc.ec.europa.eu/

(19) Mizuno M, Yoshida M, Kodama T, Kosaka N, Okamato K, Sono S, Yamada T, Hasegawa S, Ashikaga T, Kuwahara H, Sakaguchi H, Sato J, Ota N, Okamoto Y, Ohno Y. (2008). Effects of pre-culture conditions on the human Cell Line Activation Test (h-CLAT) results; Results of the 4th Japanese inter-laboratory study. AATEX 13, 70-82.

(20) Sono S, Mizuno M, Kosaka N, Okamoto K, Kato Y, Inaba H, , Nakamura T, Kishi M, Kuwahara H, Sakaguchi H, Okamoto Y, Ashikaga T, Ohno Y. (2010). The Japanese ring study of a human Cell Line Activation Test (h-CLAT) for predicting skin sensitization potential (7th report): Evaluation of volatile, poorly soluble fragrance materials. AATEX 15, 89-96.

(21) OECD (2005). Guidance Document No 34 on The Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment. OECD Series on Testing and Assessment. Organization for Economic Cooperation and Development, Paris, France, 2005, 96 pp.

(22) OECD (2012). The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins. Part 1: Scientific Evidence. Series on Testing and Assessment No 168. Available at: http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV/JM/MONO(2012)10/PART1&docLanguage=En

(23) United Nations UN (2013). Globally Harmonized System of Classification and Labelling of Chemicals (GHS). Fifth revised edition. New York & Geneva: United Nations Publications. ISBN: 978-92-1-117006-1. Available at: http://www.unece.org/trans/danger/publi/ghs/ghs_rev05/05files_e.html

(24) ECETOC (2003). Contact sensitization: Classification according to potency. European Centre for Ecotoxicology & Toxicology of Chemicals (Technical Report No 87).

(25) Ashikaga T, Sakaguchi H, Okamoto K, Mizuno M, Sato J, Yamada T, Yoshida M, Ota N, Hasegawa S, Kodama T, Okamoto Y, Kuwahara H, Kosaka N, Sono S, Ohno Y. (2008). Assessment of the human Cell Line Activation Test (h-CLAT) for Skin Sensitization; Results of the First Japanese Inter-laboratory Study. AATEX 13, 27-35.

Appendix 1.1

DEFINITIONS

Accuracy: The closeness of agreement between test results and accepted reference values. It is a measure of test performance and one aspect of relevance. The term is often used interchangeably with concordance to mean the proportion of correct outcomes of a test (21).

AOP (Adverse Outcome Pathway): sequence of events from the chemical structure of a target chemical or group of similar chemicals through the molecular initiating event to an *in vivo* outcome of interest (22).

Chemical: A substance or a mixture.

CV75: The estimated concentration showing 75% cell viability.

EC150: the concentrations showing the RFI values of 150 in CD86 expression

EC200: the concentrations showing the RFI values of 200 in CD54 expression

Flow cytometry: a cytometric technique in which cells suspended in a fluid flow one at a time through a focus of exciting light, which is scattered in patterns characteristic to the cells and their components; cells are frequently labeled with fluorescent markers so that light is first absorbed and then emitted at altered frequencies.

Hazard: Inherent property of an agent or situation having the potential to cause adverse effects when an organism, system or (sub) population is exposed to that agent.

IATA (Integrated Approach to Testing and Assessment): A structured approach used for hazard identification (potential), hazard characterisation (potency) and/or safety assessment (potential/potency and exposure) of a chemical or group of chemicals, which strategically integrates and weights all relevant data to inform regulatory decision regarding potential hazard and/or risk and/or the need for further targeted and therefore minimal testing.

Medium control: An untreated replicate containing all components of a test system. This sample is processed with test chemical-treated samples and other control samples to determine whether the solvent/vehicle interacts with the test system.

Mixture: A mixture or a solution composed of two or more substances.

Mono-constituent substance: A substance, defined by its quantitative composition, in which one main constituent is present to at least 80% (w/w).

Multi-constituent substance: A substance, defined by its quantitative composition, in which more than one main constituent is present in a concentration ≥ 10% (w/w) and < 80% (w/w). A multi-constituent substance is the result of a manufacturing process. The difference between mixture and multi-constituent substance is that a mixture is obtained by

blending of two or more substances without chemical reaction. A multi-constituent substance is the result of a chemical reaction.

Positive control: A replicate containing all components of a test system and treated with a substance known to induce a positive response. To ensure that variability in the positive control response across time can be assessed, the magnitude of the positive response should not be excessive.

Pre-haptens: chemicals which become sensitisers through abiotic transformation

Pro-haptens: chemicals requiring enzymatic activation to exert skin sensitisation potential

Relative fluorescence intensity (RFI): Relative values of geometric mean fluorescence intensity (MFI) in chemical-treated cells compared to MFI in solvent/vehicle-treated cells.

Relevance: Description of relationship of the test to the effect of interest and whether it is meaningful and useful for a particular purpose. It is the extent to which the test correctly measures or predicts the biological effect of interest. Relevance incorporates consideration of the accuracy (concordance) of a test (21).

Reliability: Measures of the extent that a test can be performed reproducibly within and between laboratories over time, when performed using the same protocol. It is assessed by calculating intra- and inter-laboratory reproducibility and intra-laboratory repeatability (21).

Run: A run consists of one or more test chemicals tested concurrently with a solvent/vehicle control and with a positive control.

Sensitivity: The proportion of all positive/active chemicals that are correctly classified by the test. It is a measure of accuracy for a test that produces categorical results, and is an important consideration in assessing the relevance of a test(21).

Staining buffer: A phosphate buffered saline containing 0.1% bovine serum albumin.

Solvent/vehicle control: An untreated sample containing all components of a test system except of the test chemical, but including the solvent/vehicle that is used. It is used to establish the baseline response for the samples treated with the test chemical dissolved or stably dispersed in the same solvent/vehicle. When tested with a concurrent medium control, this sample also demonstrates whether the solvent/vehicle interacts with the test system.

Specificity: The proportion of all negative/inactive chemicals that are correctly classified by the test. It is a measure of accuracy for a test that produces categorical results and is an important consideration in assessing the relevance of a test (21).

Substance: A chemical element and its compounds in the natural state or obtained by any production process, inducing any additive necessary to preserve its stability and any impurities deriving from the process used, but excluding any solvent which may be separated without affecting the stability of the substance or changing it composition.

Test chemical: Any substance or mixture tested using this method.

United Nations Globally Harmonized System of Classification and Labelling of Chemicals (UN GHS): A system proposing the classification of chemicals (substances and mixtures) according to standardised types and levels of physical, health and environmental hazards, and addressing corresponding communication elements, such as pictograms, signal words, hazard statements, precautionary statements and safety data sheets, so that to convey information on their adverse effects with a view to protect people (including employers, workers, transporters, consumers and emergency responders) and the environment (23).

UVCB: substances of unknown or variable composition, complex reaction products or biological materials.

Valid test: A test considered to have sufficient relevance and reliability for a specific purpose and which is based on scientifically sound principles. A test is never valid in an absolute sense, but only in relation to a defined purpose (21).

Appendix 1.2

PROFICIENCY SUBSTANCES

Prior to routine use of the test described in this Appendix to test method B.71, laboratories should demonstrate technical proficiency by correctly obtaining the expected h-CLAT prediction for the 10 substances recommended in Table 1 and by obtaining CV75, EC150 and EC200 values that fall within the respective reference range for at least 8 out of the 10 proficiency substances. Proficiency substances were selected to represent the range of responses for skin sensitisation hazards. Other selection criteria were that the substances are commercially available, and that high-quality *in vivo* reference data as well as high quality *in vitro* data generated with the h-CLAT method are available. Also, published reference data are available for the h-CLAT method (3) (14).

Table 1: Recommended substances for demonstrating technical proficiency with the h-CLAT method

Proficiency substances	**CASRN**	**Physical state**	***In vivo* prediction[1]**	**CV75 Reference Range in µg/ml[2]**	**h-CLAT results for CD86 (EC150 Reference Range in µg/ml)[2]**	**h-CLAT results for CD54 (EC200 Reference Range in µg/ml)[2]**
2,4-Dinitrochlorobenzene	97-00-7	Solid	Sensitiser (extreme)	2-12	Positive (0.5-10)	Positive (0.5-15)
4-Phenylenediamine	106-50-3	Solid	Sensitiser (strong)	5-95	Positive (<40)	Negative (>1.5)[3]
Nickel sulfate	10101-97-0	Solid	Sensitiser (moderate)	30-500	Positive (<100)	Positive (10-100)
2-Mercaptbenzothiazole	149-30-4	Solid	Sensitiser (moderate)	30-400	Negative (>10)[3]	Positive (10-140)
R(+)-Limonene	5989-27-5	Liquid	Sensitiser (weak)	>20	Negative (>5)[3]	Positive (<250)
Imidazolidinyl urea	39236-46-9	Solid	Sensitiser (weak)	25-100	Positive (20-90)	Positive (20-75)
Isopropanol	67-63-0	Liquid	Non-sensitiser	>5000	Negative (>5000)	Negative (>5000)
Glycerol	56-81-5	Liquid	Non-sensitiser	>5000	Negative (>5000)	Negative (>5000)
Lactic acid	50-21-5	Liquid	Non-sensitiser	1500-5000	Negative (>5000)	Negative (>5000)
4-Aminobenzoic acid	150-13-0	Solid	Non-sensitiser	>1000	Negative (>1000)	Negative (>1000)

Abbreviations: CAS RN = Chemical Abstracts Service Registry Number

[1] The *in vivo* hazard and (potency) prediction is based on LLNA data (3) (14). The *in vivo* potency is derived using the criteria proposed by ECETOC (24).

[2] Based on historical observed values (13) (25).

[3] Historically, a majority of negative results have been obtained for this marker and therefore a negative result is mostly expected. The range provided was defined on the basis of the few historical positive results observed. In case a positive result is obtained, the EC value should be within the reported reference range.

Appendix 2

IN VITRO SKIN SENSITISATION: U937 CELL LINE ACTIVATION TEST (U-SENS™)

INITIAL CONSIDERATIONS AND LIMITATIONS

1. The U-SENS™ test quantifies the change in the expression of a cell surface marker associated with the process of activation of monocytes and dendritic cells (DC) (i.e. CD86), in the human histiocytic lymphoma cell line U937, following exposure to sensitisers (1). The measured expression levels of CD86 cell surface marker in the cell line U937 is then used for supporting the discrimination between skin sensitisers and non-sensitisers.

2. The U-SENS™ test has been evaluated in a validation study (2) coordinated by L'Oreal and subsequently independent peer reviewed by the European Union Reference Laboratory for Alternatives to Animal Testing (EURL ECVAM) Scientific Advisory Committee (ESAC) (3). Considering all available evidence and input from regulators and stakeholders, the U-SENS™ was recommended by EURL ECVAM (4) to be used as part of an IATA to support the discrimination between sensitisers and non-sensitisers for the purpose of hazard classification and labelling. In its guidance document on the reporting of structured approaches to data integration and individual information sources used within IATA for skin sensitisation, the OECD currently discusses a number of case studies describing different testing strategies and prediction models. One of the different defined approaches is based on the U-SENS assay (5). Examples of the use of U-SENS™ data in combination with other information, including historical data and existing valid human data (6), are also reported elsewhere in the literature (4) (5) (7).

3. The U-SENS™ test proved to be transferable to laboratories experienced in cell culture techniques and flow cytometry analysis. The level of reproducibility in predictions that can be expected from the test is in the order of 90% and 84% within and between laboratories, respectively (8). Results generated in the validation study (8) and other published studies (1) overall indicate that, compared with LLNA results, the accuracy in distinguishing skin sensitisers (i.e. UN GHS/CLP Cat.1) from non-sensitisers is 86% (N=166) with a sensitivity of 91% (118/129) and a specificity of 65% (24/37). Compared with human results, the accuracy in distinguishing skin sensitisers (i.e. UN GHS/CLP Cat.1) from non-sensitisers is 77% (N=101) with a sensitivity of 100% (58/58) and a specificity of 47% (20/43). False negative predictions compared to LLNA with the U-SENS™ are more likely to concern chemicals showing a low to moderate skin sensitisation potency (i.e. UN GHS/CLP subcategory 1B) than chemicals showing a high skin sensitisation potency (i.e. UN GHS/CLP subcategory 1A) (1) (8) (9). Taken together, this information indicates the usefulness of the U-SENS™ test to contribute to the identification of skin sensitisation hazards. However, the accuracy values given here for the U-SENS™ as a stand-alone test

are only indicative, since the test should be considered in combination with other sources of information in the context of an IATA and in accordance with the provisions of paragraphs 7 and 8 in the General introduction. Furthermore, when evaluating non-animal methods for skin sensitisation, it should be kept in mind that the LLNA test as well as other animal tests may not fully reflect the situation in humans.

4. On the basis of the data currently available, the U-SENS™ test was shown to be applicable to test chemicals (including cosmetics ingredients e.g. preservatives, surfactants, actives, dyes) covering a variety of organic functional groups, of physicochemical properties, skin sensitisation potency (as determined in *in vivo* studies) and the spectrum of reaction mechanisms known to be associated with skin sensitisation (i.e. Michael acceptor, Schiff base formation, acyl transfer agent, substitution nucleophilic bi-molecular [SN2], or nucleophilic aromatic substitution [SNAr]) (1) (8) (9) (10). The U-SENS™ test is applicable to test chemicals that are soluble or that form a stable dispersion (i.e. a colloid or suspension in which the test chemical does not settle or separate from the solvent/vehicle into different phases) in an appropriate solvent/vehicle (see paragraph 13). Chemicals in the dataset reported to be pre-haptens (i.e. substances activated by oxidation) or pro-haptens (i.e. substances requiring enzymatic activation for example via P450 enzymes) were correctly predicted by the U-SENS™ (1) (10). Membrane disrupting substances can lead to false positive results due to a non-specific increase of CD86 expression, as 3 out of 7 false positives relative to the *in vivo* reference classification were surfactants (1). As such positive results with surfactants should be considered with caution whereas negative results with surfactants could still be used to support the identification of the test chemical as a non-sensitiser. Fluorescent test chemicals can be assessed with the U-SENS™ (1), nevertheless, strong fluorescent test chemicals emitting at the same wavelength as fluorescein isothiocyanate (FITC) or as propidium iodide (PI), will interfere with the flow cytometric detection and thus cannot be correctly evaluated using FITC-conjugated antibodies (potential false negative) or PI (viability not measurable). In such a case, other fluorochrome-tagged antibodies or other cytotoxicity markers, respectively, can be used as long as it can be shown they provide similar results as the FITC-tagged antibodies or PI (see paragraph 18) e.g. by testing the proficiency substances in Appendix 2.2. In the light of the above, positive results with surfactants and negative results with strong fluorescent test chemicals should be interpreted in the context of the stated limitations and together with other information sources within the framework of IATA. In cases where there is evidence demonstrating the non-applicability of the U-SENS™ test to other specific categories of test chemicals, it should not be used for those specific categories.

5. As described above, the U-SENS™ test supports the discrimination between skin sensitisers from non-sensitisers. However, it may also potentially contribute to the assessment of sensitising potency when used in integrated approaches such as IATA.

Nevertheless, further work, preferably based on human data, is required to determine how U-SENS™ results may possibly inform potency assessment.

6. Definitions are provided in Appendix 2.1.

PRINCIPLE OF THE TEST

7. The U-SENS™ test is an *in vitro* assay that quantifies changes of CD86 cell surface marker expression on a human histiocytic lymphoma cell line, U937 cells, following 45±3 hours exposure to the test chemical. The CD86 surface marker is one typical marker of U937 activation. CD86 is known to be a co-stimulatory molecule that may mimic monocytic activation, which plays a critical role in T-cell priming. The changes of CD86 cell surface marker expression are measured by flow cytometry following cell staining typically with fluorescein isothiocyanate (FITC)-labelled antibodies. Cytotoxicity measurement is also conducted (e.g. by using PI) concurrently to assess whether upregulation of CD86 cell surface marker expression occurs at sub-cytotoxic concentrations. The stimulation index (S.I.) of CD86 cell surface marker compared to solvent/vehicle control is calculated and used in the prediction model (see paragraph 19), to support the discrimination between sensitisers and non-sensitisers.

DEMONSTRATION OF PROFICIENCY

8. Prior to routine use of the test described in this Appendix to test method B.71, laboratories should demonstrate technical proficiency, using the 10 Proficiency Substances listed in Appendix 2.2 in compliance with the Good *in vitro* Method Practices (11). Moreover, test users should maintain a historical database of data generated with the reactivity checks (see paragraph 11) and with the positive and solvent/vehicle controls (see paragraphs 15-16), and use these data to confirm the reproducibility of the test in their laboratory is maintained over time.

PROCEDURE

9. This test is based on the U-SENS™ DataBase service on ALternative Methods to animal experimentation (DB-ALM) protocol no 183 (12). The Standard Operating Procedures (SOP) should be employed when implementing and using the U-SENS™ test in the laboratory. An automated system to run the U-SENS™ can be used if it can be shown to provide similar results, for example by testing the proficiency substances in Appendix 2.2. The following is a description of the main components and procedures for the U-SENS™ test.

Preparation of cells

10. The human histiocytic lymphoma cell line, U937 (13) should be used for performing the U-SENS™ test. Cells (clone CRL1593.2) should be obtained from a well-qualified cell bank such as the American Type Culture Collection.

11. U937 cells are cultured, at 37°C under 5% CO_2 and humidified atmosphere, in RPMI-1640 medium supplemented with 10% foetal calf serum (FCS), 2 mM L-glutamine, 100 units/ml penicillin and 100 µg/ml streptomycin (complete medium). U937 cells are routinely passaged every 2-3 days at the density of 1.5 or 3 × 10^5 cells/ml, respectively. The cell density should not exceed 2 × 10^6 cells/ml and the cell viability measured by trypan blue exclusion should be ≥ 90% (not to be applied at the first passage after thawing). Prior to using them for testing, every batch of cells, FCS or antibodies should be qualified by conducting a reactivity check. The reactivity check of the cells should be performed using the positive control, picrylsulfonic acid (2,4,6-Trinitro-benzene-sulfonic acid: TNBS) (CASRN 2508-19-2, ≥ 99% purity) and the negative control lactic acid (LA) (CASRN 50-21-5, ≥ 85% purity), at least one week after thawing. For the reactivity check, six final concentrations should be tested for each of the 2 controls (TNBS: 1, 12.5, 25, 50, 75, 100µg/ml and LA: 1, 10, 20, 50, 100, 200µg/ml). TNBS solubilised in complete medium should produce a positive and concentration-related response of CD86 (e.g. when a positive concentration, CD86 S.I. ≥ 150, is followed by a concentration with an increasing CD86 S.I), and LA solubilised in complete medium should produce negative response of CD86 (see paragraph 21). Only the batch of cells which passed the reactivity check 2 times should be used for the assay. Cells can be propagated up to seven weeks after thawing. Passage number should not exceed 21. The reactivity check should be performed according to the procedures described in paragraphs 18-22.

12. For testing, U937 cells are seeded at a density of either 3 x 10^5 cells/ml or 6 × 10^5 cells/ml, and pre-cultured in culture flasks for 2 days or 1 day, respectively. Other pre-cultured conditions than those described above may be used if sufficient scientific rationale is provided and if it can be shown to provide similar results, for example by testing the proficiency substances in Appendix 2.2. In the day of testing, cells harvested from culture flask are resuspended with fresh culture medium at 5 × 10^5 cells/ml. Then, cells are distributed into a 96-well flat-bottom plate with 100 µl (final cell density of 0.5 × 10^5 cells/well).

Preparation of test chemicals and control substances

13. Assessment of solubility is conducted prior to testing. For this purpose, test chemicals are dissolved or stably dispersed at a concentration of 50 mg/ml in complete medium as first solvent option or dimethyl sulfoxide (DMSO, ≥ 99% purity) as a second solvent/vehicle option if the test chemical is not soluble in the complete medium solvent/vehicle. For the testing, the test chemical is dissolved to a final concentration of 0.4 mg/ml in complete medium if the chemical is soluble in this solvent/vehicle. If the chemical is soluble only in

DMSO, the chemical is dissolved at a concentration of 50 mg/ml. Other solvents/vehicles than those described above may be used if sufficient scientific rationale is provided. Stability of the test chemical in the final solvent/vehicle should be taken into account.

14. The test chemicals and control substances are prepared on the day of testing. Because a dose finding assay is not conducted, for the first run, 6 final concentrations should be tested (1, 10, 20, 50, 100 and 200 µg/ml) into the corresponding solvent/vehicle either in complete medium or in 0.4% DMSO in medium. For the subsequent runs, starting from the 0.4 mg/ml in complete medium or 50 mg/ml in DMSO, solutions of the test chemicals, at least 4 working solutions (i.e. at least 4 concentrations), are prepared using the corresponding solvent/vehicle. The working solutions are finally used for treatment by adding an equal volume of U937 cell suspension (see paragraph 11 above) to the volume of working solution in the plate to achieve a further 2-fold dilution (12). The concentrations (at least 4 concentrations) for any further run are chosen based on the individual results of all previous runs (8). The usable final concentrations are 1, 2, 3, 4, 5, 7.5, 10, 12.5, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 120, 140, 160, 180 and 200 µg/ml. The maximum final concentration is 200 µg/ml. In the case of a CD86 positive value at 1 µg/ml is observed, then 0.1 µg/ml is evaluated in order to find the concentration of the test chemical that does not induce CD86 above the positive threshold. For each run, the EC150 (concentration at which a chemical reaches the CD86 positive threshold of 150%, see paragraph 19) is calculated if a CD86 positive concentration-response is observed. Where the test chemical induces a positive CD86 response not concentration related, the calculation of the EC150 might not be relevant as described in the U-SENS™ DB-ALM protocol no 183 (12). For each run, CV70 (concentration at which a chemical reaches the cytotoxicity threshold of 70%, see paragraph 19) is calculated whenever possible (12). To investigate the concentration response effect of CD86 increase, any concentrations from the usable concentrations should be chosen evenly spread between the EC150 (or the highest CD86 negative non cytotoxic concentration) and the CV70 (or the highest concentration allowed i.e. 200 µg/ml). A minimum of 4 concentrations should be tested per run with at least 2 concentrations being common with the previous run(s), for comparison purposes.

15. The solvent/vehicle control used in the U-SENS™ test is complete medium (for test chemicals solubilised or stably dispersed) (see paragraph 4) or 0.4% DMSO in complete medium (for test chemicals solubilised or stably dispersed in DMSO).

16. The positive control used in the U-SENS™ test is TNBS (see paragraph 11), prepared in complete medium. TNBS should be used as the positive control for CD86 expression measurement at a final single concentration in plate (50 µg/ml) yielding > 70% of cell viability. To obtain a 50 µg/ml concentration of TNBS in plate, a 1 M (i.e. 293 mg/ml) stock solution of TNBS in complete medium is prepared and further diluted 2930-fold with complete medium to a 100 µg/ml working solution. Lactic acid (LA, CAS 50-21-5) should

be used as the negative control at 200 µg/ml solubilised in complete medium (from a 0.4 mg/ml stock solution). In each plate of each run, three replicates of complete medium untreated control, solvent/vehicle control, negative and positive controls are prepared (12). Other suitable positive controls may be used if historical data are available to derive comparable run acceptance criteria. The run acceptance criteria are the same as described for the test chemical (see paragraph 12).

Application of test chemicals and control substances

17. The solvent/vehicle control or working solutions described in paragraphs 14-16 are mixed 1:1 (v/v) with the cell suspensions prepared in the 96-well flat-bottom plate (see paragraph 12). The treated plates are then incubated for 45±3 hours at 37°C under 5% CO_2. Prior to incubation, plates are sealed with semi permeable membrane, to avoid evaporation of volatile test chemicals and cross-contamination between cells treated with test chemicals (12).

Cell staining

18. After 45±3 hours of exposure, cells are transferred into V-shaped microtiter plate and collected by centrifugation. Solubility interference is defined as crystals or drops observed under the microscope at 45 ± 3 hours post treatment (before the cell staining). The supernatants are discarded and the remaining cells are washed once with 100 µl of an ice-cold phosphate buffered saline (PBS) containing 5 % foetal calf serum (staining buffer). After centrifugation, cells are re-suspended with 100 µl of staining buffer and stained with 5 µl (e.g. 0.25 µg) of FITC-labelled anti-CD86 or mouse IgG1 (isotype) antibodies at 4°C for 30 min protected from light. The antibodies described in the U-SENS™ DB-ALM protocol no 183 (12) should be used (for CD86: BD-PharMingen #555657 Clone: Fun-1, or Caltag/Invitrogen # MHCD8601 Clone: BU63; and for IgG1: BD-PharMingen #555748, or Caltag/Invitrogen # GM4992). Based on the experience of the test developers, the fluorescence intensity of the antibodies is usually consistent between different lots. Other clones or supplier of the antibodies which passed the reactivity check may be used for the assay (see paragraph 11). However, users may consider titrating the antibodies in their own laboratory's conditions to define the best concentration for use. Other detection system e.g. fluorochrome-tagged anti-CD86 antibodies may be used if they can be shown to provide similar results as FITC-conjugated antibodies, for example by testing the proficiency substances in Appendix 2.2. After washing with 100 µl of staining buffer two times and once with 100 µl of an ice-cold PBS, cells are resuspended in ice-cold PBS (e.g. 125 µl for samples being analysed manually tube by tube, or 50 µl using an auto-sampler plate) and PI solution is added (final concentration of 3 µg/ml). Other cytotoxicity markers, such as 7-Aminoactinomycin D (7-AAD) or Trypan blue may be used if the alternative stains can be shown to provide similar results as PI, for example by testing the proficiency substances in Appendix 2.2.

Flow cytometry analysis

19. Expression level of CD86 and cell viability are analysed using flow cytometry. Cells are displayed within a size (FSC) and granularity (SSC) dot plot set to log scale in order to clearly identify the population in a first gate R1 and eliminate the debris. A targeting total of 10 000 cells in gate R1 are acquired for each well. Cells from the same R1 gate are displayed within a FL3 or FL4 / SSC dot plot. Viable cells are delineated by placing a second gate R2 selecting the population of propidium iodide-negative cells (FL3 or FL4 channel). The cell viability can be calculated using the following equation by the cytometer analysis program. When the cell viability is low, up to 20 000 cells including dead cells could be acquired. Alternatively, data can be acquired for one minute after the initiation of the analysis.

$$Cell\ viability = \frac{Number\ of\ living\ cells}{Total\ Number\ of\ aquired\ cells} \times 100$$

Percentage of FL1-positive cells is then measured among these viable cells gated on R2 (within R1). Cell surface expression of CD86 is analysed in a FL1 / SSC dot plot gated on viable cells (R2).

For the complete medium / IgG1 wells, the analysis marker is set close to the main population so that the complete medium controls have IgG1 within the target zone of 0.6 to 0.9%.

Colour interference is defined as a shift of the FITC-labelled IgG1 dot-plot (IgG1 FL1 Geo Mean S.I. ≥ 150%).

The stimulation index (S.I.) of CD86 for controls cells (untreated or in 0.4% DMSO) and chemical-treated cells are calculated according to the following equation:

$$S.I. = \frac{\%\ of\ CD86^{+} treated\ cells - \%\ of\ IgG1^{+} treated\ cells}{\%\ of\ CD86^{+}\ control\ cells - \%\ of\ IgG1^{+}\ control\ cells} \times 100$$

% of $IgG1^{+}$ untreated control cells: referred to as percentage of FL1-positive IgG1 cells defined with the analysis marker (accepted range of ≥ 0.6% and < 1.5%, see paragraph 22) among the viable untreated cells.

% of $IgG1^{+}/CD86^{+}$ control/treated cells: referred to as percentage of FL1-positive IgG1/CD86 cells measured without moving the analysis marker among the viable control/treated cells.

DATA AND REPORTING

Data evaluation

20. The following parameters are calculated in the U-SENS™ test: CV70 value, i.e. a concentration showing 70% of U937 cell survival (30% cytotoxicity) and the EC150 value,

i.e. the concentration at which the test chemicals induced a CD86 stimulation index (S.I.) of 150%.

CV70 is calculated by log-linear interpolation using the following equation:

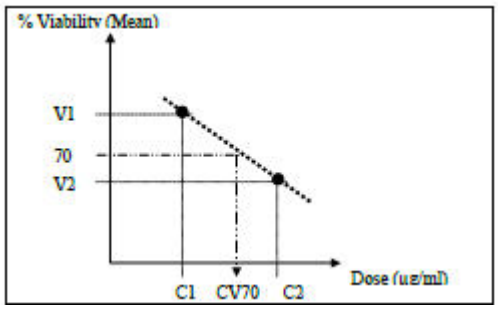
CV70 = C1 + [(V1 - 70) / (V1 – V2) * (C2 – C1)]

Where:

V1 is the minimum value of cell viability over 70%

V2 is the maximum value of cell viability below 70%

C1 and C2 are the concentrations showing the value of cell viability V1 and V2 respectively.



Other approaches to derive the CV70 can be used as long as it is demonstrated that this has no impact on the results (e.g. by testing the proficiency substances).

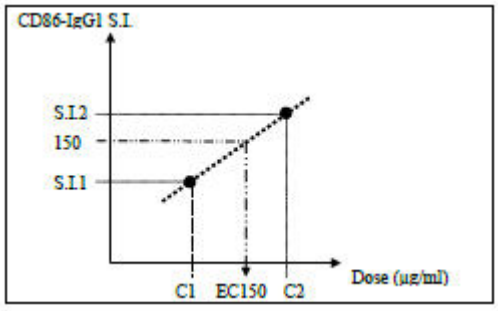
EC150 is calculated by log-linear interpolation using the following equation:

EC150 = C1 + [(150 – S.I.1) / (S.I.2 – S.I.1) * (C2 – C1)]

Where:

C1 is the highest concentration in µg/ml with a CD86 S.I. < 150% (S.I. 1)

C2 is the lowest concentration in µg/ml with a CD86 S.I. ≥ 150% (S.I. 2).



The EC150 and CV70 values are calculated

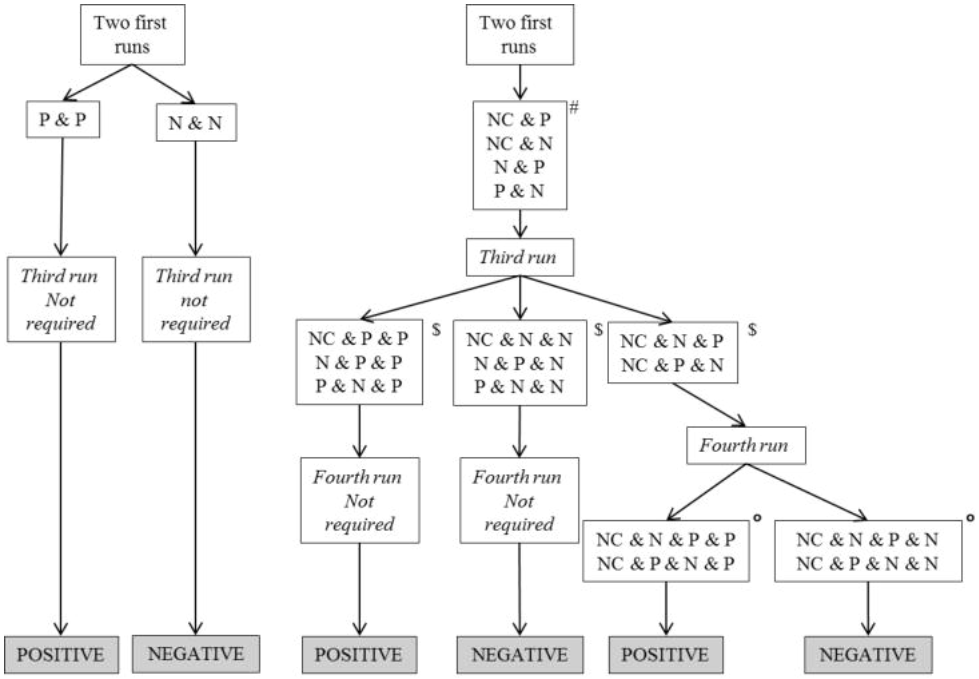
- for each run: the individual EC150 and CV70 values are used as tools to investigate the concentration response effect of CD86 increase (see paragraph 14),
- based on the average viabilities, the overall CV70 is determined (12) ,
- based on the average S.I. of CD86 values, the overall EC150 is determined for the test chemical predicted as POSITIVE with the U-SENS™ (see paragraph 21) (12).

Prediction model

21. For CD86 expression measurement, each test chemical is tested in at least four concentrations and in at least two independent runs (performed on a different day) to derive a single prediction (NEGATIVE or POSITIVE).

- The individual conclusion of an U-SENS™ run is considered Negative (hereinafter referred to as N) if the S.I. of CD86 is less than 150% at all non-cytotoxic concentrations (cell viability ≥ 70%) and if no interference is observed (cytotoxicity, solubility: see paragraph 18 or colour: see paragraph 19 regardless of the non-cytotoxic concentrations at which the interference is detected). In all other cases: S.I. of CD86 higher or equal to 150% and/or interferences observed, the individual conclusion of an U-SENS™ run is considered Positive (hereinafter referred to as P).
- An U-SENS™ prediction is considered NEGATIVE if at least two independent runs are negative (N) (Figure 1). If the first two runs are both negative (N), the U-SENS™ prediction is considered NEGATIVE and a third run does not need to be conducted.
- An U-SENS™ prediction is considered POSITIVE if at least two independent runs are positive (P) (Figure 1). If the first two runs are both positive (P), the U-SENS™ prediction is considered POSITIVE and a third run does not need to be conducted.
- Because a dose finding assay is not conducted, there is an exception if, in the first run, the S.I. of CD86 is higher or equal to 150% at the highest non-cytotoxic concentration only. The run is then considered to be NOT CONCLUSIVE (NC), and additional concentrations (between the highest non cytotoxicity concentration and the lowest cytotoxicity concentration - see paragraph 20) should be tested in additional runs. In case a run is identified as NC, at least 2 additional runs should be conducted, and a fourth run in case runs 2 and 3 are not concordant (N and/or P independently) (Figure 1). Follow up runs will be considered positive even if only one non cytotoxic concentration gives a CD86 equal or above 150%, since the concentration setting has been adjusted for the specific test chemical. The final prediction will be based on the majority result of the three or four individual runs (i.e. 2 out of 3 or 2 out of 4) (Figure 1).

Figure 1: Prediction model used in the U-SENS™ test. An U-SENS™ prediction should be considered in the framework of an IATA and in accordance with the provision of paragraph 4 and of the General introduction paragraphs 7, 8 and 9.



N: Run with no CD86 positive or interference observed;

P: Run with CD86 positive and/or interference(s) observed;

NC: Not Conclusive. First run with No Conclusion when CD86 is positive at the highest non-cytotoxic concentration only;

#: A Not Conclusive (NC) individual conclusion attributed only to the first run conducts automatically to the need of a third run to reach a majority of Positive (P) or Negative (N) conclusions in at least 2 of 3 independent runs.

$: The boxes show the relevant combinations of results from the three runs on the basis of the results obtained in the first two runs shown in the box above.

°: The boxes show the relevant combinations of results from the four runs on the basis of the results obtained in the first three runs shown in the box above.

Acceptance criteria

22. The following acceptance criteria should be met when using the U-SENS™ test (12).

- At the end of the 45±3 hours exposure period, the mean viability of the triplicate untreated U937 cells had to be > 90% and no drift in CD86 expression is observed. The CD86 basal expression of untreated U937 cells had to be comprised within the range of ≥ 2% and ≤ 25%.
- When DMSO is used as a solvent, the validity of the DMSO vehicle control is assessed by calculating a DMSO S.I. compared to untreated cells, and the mean viability of the triplicate cells had to be > 90%. The DMSO vehicle control is valid if the mean value of its triplicate CD86 S.I. was smaller than 250% of the mean of the triplicate CD86 S.I. of untreated U937 cells.

- The runs are considered valid if at least two out of three IgG1 values of untreated U937 cells fell within the range of ≥ 0.6% and < 1.5%.
- The concurrent tested negative control (lactic acid) is considered valid if at least two out of the three replicates were negative (CD86 S.I. < 150%) and non-cytotoxic (cell viability ≥ 70%).
- The positive control (TNBS) was considered as valid if at least two out of the three replicates were positive (CD86 S.I. ≥ 150%) and non-cytotoxic (cell viability ≥ 70%).

Test report

23. The test report should include the following information.

Test Chemical

Mono-constituent substance

- Chemical identification, such as IUPAC or CAS name(s), CAS number(s), SMILES or InChI code, structural formula, and/or other identifiers;
- Physical appearance, complete medium solubility, DMSO solubility, molecular weight, and additional relevant physicochemical properties, to the extent available;
- Purity, chemical identity of impurities as appropriate and practically feasible, etc.;
- Treatment prior to testing, if applicable (e.g. warming, grinding);
- Concentration(s) tested;
- Storage conditions and stability to the extent available;
- Justification for choice of solvent/vehicle for each test chemical.

Multi-constituent substance, UVCB and mixture:

- Characterisation as far as possible by e.g. chemical identity (see above), purity, quantitative occurrence and relevant physicochemical properties (see above) of the constituents, to the extent available;
- Physical appearance, complete medium solubility, DMSO solubility and additional relevant physicochemical properties, to the extent available;
- Molecular weight or apparent molecular weight in case of mixtures/polymers of known compositions or other information relevant for the conduct of the study;
- Treatment prior to testing, if applicable (e.g. warming, grinding);
- Concentration(s) tested;
- Storage conditions and stability to the extent available;

- Justification for choice of solvent/vehicle for each test chemical.

Controls

Positive control

- Chemical identification, such as IUPAC or CAS name(s), CAS number(s), SMILES or InChI code, structural formula, and/or other identifiers;
- Physical appearance, DMSO solubility, molecular weight, and additional relevant physicochemical properties, to the extent available and where applicable;
- Purity, chemical identity of impurities as appropriate and practically feasible, etc.;
- Treatment prior to testing, if applicable (e.g. warming, grinding);
- Concentration(s) tested;
- Storage conditions and stability to the extent available;
- Reference to historical positive control results demonstrating suitable run acceptance criteria, if applicable.

Negative and solvent/vehicle control

- Chemical identification, such as IUPAC or CAS name(s), CAS number(s), SMILES or InChI code, structural formula, and/or other identifiers;
- Purity, chemical identity of impurities as appropriate and practically feasible, etc.;
- Physical appearance, molecular weight, and additional relevant physicochemical properties in the case other control solvent/vehicle than those mentioned in the Test Guideline are used and to the extent available;
- Storage conditions and stability to the extent available;
- Justification for choice of solvent/vehicle for each test chemical.

Test Conditions

- Name and address of the sponsor, test facility and study director;
- Description of test used;
- Cell line used, its storage conditions and source (e.g. the facility from which they were obtained);
- Flow cytometry used (e.g. model), including instrument settings, antibodies and cytotoxicity marker used;
- The procedure used to demonstrate proficiency of the laboratory in performing the test by testing of proficiency substances, and the procedure used to demonstrate reproducible

performance of the test over time, e.g. historical control data and/or historical reactivity checks' data.

Test Acceptance Criteria

- Cell viability and CD86 S.I values obtained with the solvent/vehicle control in comparison to the acceptance ranges;
- Cell viability and S.I. values obtained with the positive control in comparison to the acceptance ranges;
- Cell viability of all tested concentrations of the tested chemical.

Test procedure

- Number of runs used;
- Test chemical concentrations, application and exposure time used (if different than the one recommended)
- Duration of exposure;
- Description of evaluation and decision criteria used;
- Description of any modifications of the test procedure.

Results

- Tabulation of the data, including CV70 (if applicable), S.I., cell viability values, EC150 values (if applicable) obtained for the test chemical and for the positive control in each run, and an indication of the rating of the test chemical according to the prediction model;
- Description of any other relevant observations, if applicable.

Discussion of the Results

- Discussion of the results obtained with the U-SENS™ test;
- Consideration of the test results within the context of an IATA, if other relevant information is available.

Conclusions

LITERATURE

(1) Piroird, C., Ovigne, J.M., Rousset, F., Martinozzi-Teissier, S., Gomes, C., Cotovio, J., Alépée, N. (2015). The Myeloid U937 Skin Sensitization Test (U-SENS) addresses the activation of dendritic cell event in the adverse outcome pathway for skin sensitization. *Toxicol. In Vitro* 29, 901-916.

(2) EURL ECVAM (2017). The U-SENS™ test method Validation Study Report. Accessible at: http://ihcp.jrc.ec.europa.eu/our_labs/eurl-ecvam/eurl-ecvam-recommendations

(3) EC EURL ECVAM (2016). ESAC Opinion No 2016-03 on the L'Oréal-coordinated study on the transferability and reliability of the U-SENS™ test method for skin sensitisation testing. EUR 28178 EN; doi 10.2787/815737. Available at: [http://publications.jrc.ec.europa.eu/repository/handle/JRC103705].

(4) EC EURL ECVAM (2017). EURL ECVAM Recommendation on the use of non-animal approaches for skin sensitisation testing. EUR 28553 EN; doi 10.2760/588955. Available at: https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/eurl-ecvam-recommendation-use-non-animal-approaches-skin-sensitisation-testing.

(5) Steiling, W. (2016). Safety Evaluation of Cosmetic Ingredients Regarding their Skin Sensitization Potential. doi:10.3390/cosmetics3020014.*Cosmetics* 3, 14.

(6) OECD (2016). Guidance Document on The Reporting of Defined Approaches and Individual Information Sources to be Used Within Integrated Approaches to Testing and Assessment (IATA) For Skin Sensitisation, Series on Testing & Assessment No 256, ENV/JM/MONO(2016)29. Organisation for Economic Cooperation and Development, Paris. Available at: [http://www.oecd.org/env/ehs/testing/series-testing-assessment-publications-number.htm.

(7) Urbisch, D., Mehling, A., Guth, K., Ramirez, T., Honarvar, N., Kolle, S., Landsiedel, R., Jaworska, J., Kern, P.S., Gerberick, F., Natsch, A., Emter, R., Ashikaga, T., Miyazawa, M., Sakaguchi, H. (2015). Assessing skin sensitization hazard in mice and men using non-animal test methods. *Regul. Toxicol. Pharmacol.* 71, 337-351.

(8) Alépée, N., Piroird, C., Aujoulat, M., Dreyfuss, S., Hoffmann, S., Hohenstein, A., Meloni, M., Nardelli, L., Gerbeix, C., Cotovio, J. (2015). Prospective multicentre study of the U-SENS test method for skin sensitization testing. *Toxicol In Vitro* 30, 373-382.

(9) Reisinger, K., Hoffmann, S., Alépée, N., Ashikaga, T., Barroso, J., Elcombe, C., Gellatly, N., Galbiati, V., Gibbs, S., Groux, H., Hibatallah, J., Keller, D., Kern, P., Klaric, M., Kolle, S., Kuehnl, J., Lambrechts, N., Lindstedt, M., Millet, M., Martinozzi-Teissier, S., Natsch, A., Petersohn, D., Pike, I., Sakaguchi, H., Schepky, A., Tailhardat, M., Templier, M., van Vliet, E., Maxwell, G. (2014). Systematic evaluation of non-animal test methods for skin sensitisation safety assessment. *Toxicol. In Vitro* 29, 259-270.

(10) Fabian, E., Vogel, D., Blatz, V., Ramirez, T., Kolle, S., Eltze, T., van Ravenzwaay, B., Oesch, F., Landsiedel, R. (2013). Xenobiotic metabolizin enzyme activities in cells used for testing skin sensitization *in vitro*. *Arch. Toxicol.* 87, 1683-1696.

(11) OECD. (2018). Draft Guidance document: Good *In Vitro* Method Practices (GIVIMP) for the Development and Implementation of *In Vitro* Methods for Regulatory Use in Human Safety Assessment. Organisation for Economic Cooperation and Development, Paris. Available at: http://www.oecd.org/env/ehs/testing/OECD Final Draft GIVIMP.pdf.

(12) DB-ALM (2016). Protocol no 183: Myeloid U937 Skin Sensitization Test (U-SENS™), 33pp. Accessible at: [http://ecvam-dbalm.jrc.ec.europa.eu/].

(13) Sundström, C., Nilsson, K. (1976). Establishment and characterization of a human histiocytic lymphoma cell line (U-937). *Int. J. Cancer* 17, 565-577.

(14) OECD (2005). Series on Testing and Assessment No. 34: Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment. Organisation for Economic Cooperation and Development, Paris. Available at: http://www.oecd.org/env/ehs/testing/series-testing-assessment-publications-number.htm.

(15) United Nations UN (2015). Globally Harmonized System of Classification and Labelling of Chemicals (GHS). ST/SG/AC.10/30/Rev.6, Sixth Revised Edition, New York & Geneva: United Nations Publications. Available at: http://www.unece.org/fileadmin/DAM/trans/danger/publi/ghs/ghs_rev06/English/ST-SG-AC10-30-Rev6e.pdf.

(16) OECD (2012). Series on Testing and Assessment No 168: The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins. Part 1: Scientific Evidence. Organisation for Economic Cooperation and Development, Paris. Available at: http://www.oecd.org/env/ehs/testing/series-testing-assessment-publications-number.htm.

(17) ECETOC (2003). Technical Report No 87: Contact sensitization: Classification according to potency. European Centre for Ecotoxicology & Toxicology of Chemicals, Brussels. Available at:

https://ftp.cdc.gov/pub/Documents/OEL/06.%20Dotson/References/ECETOC_2003-TR87.pdf.

Appendix 2.1

DEFINITIONS

Accuracy: The closeness of agreement between test results and accepted reference values. It is a measure of test performance and one aspect of relevance. The term is often used interchangeably with concordance to mean the proportion of correct outcomes of a test (14).

AOP (Adverse Outcome Pathway): sequence of events from the chemical structure of a target chemical or group of similar chemicals through the molecular initiating event to an *in vivo* outcome of interest (15).

CD86 Concentration response: There is concentration-dependency (or concentration response) when a positive concentration (CD86 S.I. ≥ 150) is followed by a concentration with an increasing CD86 S.I.

Chemical: A substance or a mixture.

CV70: The estimated concentration showing 70% cell viability.

Drift: A drift is defined by i) the corrected %$CD86^+$ value of the untreated control replicate 3 is less than 50% of the mean of the corrected %$CD86^+$ value of untreated control replicates 1and 2; and ii) the corrected %$CD86^+$ value of the negative control replicate 3 is less than 50% of mean of the corrected %$CD86^+$ value of negative control replicates 1and 2.

EC150: the estimated concentrations showing the 150% S.I. of CD86 expression.

Flow cytometry: a cytometric technique in which cells suspended in a fluid flow one at a time through a focus of exciting light, which is scattered in patterns characteristic to the cells and their components; cells are frequently labeled with fluorescent markers so that light is first absorbed and then emitted at altered frequencies.

Hazard: Inherent property of an agent or situation having the potential to cause adverse effects when an organism, system or (sub) population is exposed to that agent.

IATA (Integrated Approach to Testing and Assessment): A structured approach used for hazard identification (potential), hazard characterisation (potency) and/or safety assessment (potential/potency and exposure) of a chemical or group of chemicals, which strategically integrates and weights all relevant data to inform regulatory decision regarding potential hazard and/or risk and/or the need for further targeted and therefore minimal testing.

Mixture: A mixture or a solution composed of two or more substances.

Mono-constituent substance: A substance, defined by its quantitative composition, in which one main constituent is present to at least 80% (w/w).

Multi-constituent substance: A substance, defined by its quantitative composition, in

which more than one main constituent is present in a concentration ≥ 10% (w/w) and < 80% (w/w). A multi-constituent substance is the result of a manufacturing process. The difference between mixture and multi-constituent substance is that a mixture is obtained by blending of two or more substances without chemical reaction. A multi-constituent substance is the result of a chemical reaction.

Positive control: A replicate containing all components of a test system and treated with a substance known to induce a positive response. To ensure that variability in the positive control response across time can be assessed, the magnitude of the positive response should not be excessive.

Pre-haptens: chemicals which become sensitisers through abiotic transformation, e.g. through oxidation.

Pro-haptens: chemicals requiring enzymatic activation to exert skin sensitisation potential.

Relevance: Description of relationship of the test to the effect of interest and whether it is meaningful and useful for a particular purpose. It is the extent to which the test correctly measures or predicts the biological effect of interest. Relevance incorporates consideration of the accuracy (concordance) of a test (14).

Reliability: Measures of the extent that a test can be performed reproducibly within and between laboratories over time, when performed using the same protocol. It is assessed by calculating intra- and inter-laboratory reproducibility and intra-laboratory repeatability (14).

Run: A run consists of one or more test chemicals tested concurrently with a solvent/vehicle control and with a positive control.

Sensitivity: The proportion of all positive/active chemicals that are correctly classified by the test. It is a measure of accuracy for a test that produces categorical results, and is an important consideration in assessing the relevance of a test (14).

S.I.: Stimulation Index. Relative values of geometric mean fluorescence intensity in chemical-treated cells compared to solvent-treated cells.

Solvent/vehicle control: An untreated sample containing all components of a test system except of the test chemical, but including the solvent/vehicle that is used. It is used to establish the baseline response for the samples treated with the test chemical dissolved or stably dispersed in the same solvent/vehicle. When tested with a concurrent medium control, this sample also demonstrates whether the solvent/vehicle interacts with the test system.

Specificity: The proportion of all negative/inactive chemicals that are correctly classified by the test. It is a measure of accuracy for a test that produces categorical results and is an important consideration in assessing the relevance of a test (14).

Staining buffer: A phosphate buffered saline containing 5% foetal calf serum.

Substance: A chemical element and its compounds in the natural state or obtained by any production process, inducing any additive necessary to preserve its stability and any impurities deriving from the process used, but excluding any solvent which may be separated without affecting the stability of the substance or changing it composition.

Test chemical: Any substance or mixture tested using this test.

United Nations Globally Harmonized System of Classification and Labelling of Chemicals (UN GHS): A system proposing the classification of chemicals (substances and mixtures) according to standardized types and levels of physical, health and environmental hazards, and addressing corresponding communication elements, such as pictograms, signal words, hazard statements, precautionary statements and safety data sheets, so that to convey information on their adverse effects with a view to protect people (including employers, workers, transporters, consumers and emergency responders) and the environment (16).

UVCB: substances of unknown or variable composition, complex reaction products or biological materials.

Valid test: A test considered to have sufficient relevance and reliability for a specific purpose and which is based on scientifically sound principles. A test is never valid in an absolute sense, but only in relation to a defined purpose (14).

Appendix 2.2

PROFICIENCY SUBSTANCES

Prior to routine use of the test described in this Appendix to test method B.71, laboratories should demonstrate technical proficiency by correctly obtaining the expected U-SENS™ prediction for the 10 substances recommended in Table 1 and by obtaining CV70 and EC150 values that fall within the respective reference range for at least 8 out of the 10 proficiency substances. Proficiency substances were selected to represent the range of responses for skin sensitisation hazards. Other selection criteria were that the substances are commercially available, and that high-quality *in vivo* reference data as well as high quality *in vitro* data generated with the U-SENS™ test are available. Also, published reference data are available for the U-SENS™ test (1) (8).

Table 1: Recommended substances for demonstrating technical proficiency with the U-SENS™ test

Proficiency substances	**CASRN**	**Physical state**	***In vivo* prediction[1]**	**U-SENS™ Solvent/ Vehicle**	**U-SENS™ CV70 Reference Range in µg/ml[2]**	**U-SENS™ EC150 Reference Range in µg/ml[2]**
4-Phenylenediamine	106-50-3	Solid	Sensitiser (strong)	Complete medium[3]	<30	Positive (≤10)
Picryl sulfonic acid	2508-19-2	Liquid	Sensitizer (strong)	Complete medium	>50	Positive (≤50)
Diethyl maleate	141-05-9	Liquid	Sensitiser (moderate)	DMSO	10-100	Positive (≤20)
Resorcinol	108-46-3	Solid	Sensitiser (moderate)	Complete medium	>100	Positive (≤50)
Cinnamic alcohol	104-54-1	Solid	Sensitiser (weak)	DMSO	>100	Positive (10-100)
4-Allylanisole	140-67-0	Liquid	Sensitiser (weak)	DMSO	>100	Positive (<200)
Saccharin	81-07-2	Solid	Non-sensitiser	DMSO	>200	Negative (>200)
Glycerol	56-81-5	Liquid	Non-sensitiser	Complete medium	>200	Negative (>200)
Lactic acid	50-21-5	Liquid	Non-sensitiser	Complete medium	>200	Negative (>200)
Salicylic acid	69-72-7	Solid	Non-sensitiser	DMSO	>200	Negative (>200)

Abbreviations: CAS RN = Chemical Abstracts Service Registry Number

[1] The *in vivo* hazard and (potency) prediction is based on LLNA data (1) (8). The *in vivo* potency is derived using the criteria proposed by ECETOC (17).

[2] Based on historical observed values (1) (8).

[3] Complete medium: RPMI-1640 medium supplemented with 10% foetal calf serum, 2 mM L-glutamine, 100 units/ml

penicillin and 100 µg/ml streptomycin (8).

Appendix 3

IN VITRO SKIN SENSITISATION: IL-8 LUC ASSAY

INITIAL CONSIDERATIONS AND LIMITATIONS

1. In contrast to assays analysing the expression of cell surface markers, the IL8-Luc assay quantifies changes in IL-8 expression, a cytokine associated with the activation of dendritic cells (DC). In the THP-1-derived IL-8 reporter cell line (THP-G8, established from the human acute monocytic leukemia cell line THP-1), IL-8 expression is measured following exposure to sensitisers (1). The expression of luciferase is then used to aid discrimination between skin sensitisers and non-sensitisers.

2. The IL-8 Luc assay has been evaluated in a validation study (2) conducted by the Japanese Centre for the Validation of Alternatives Methods (JaCVAM), the **Ministry of Economy, Trade and Industry** (METI), and the Japanese Society **for Alternatives to Animal Experiments (JSAAE**) and subsequently subjected to independent peer review (3) under the auspices of JaCVAM and the Ministry of Health, Labour and Welfare (MHLW) with the support of the **International Cooperation on Alternative Test** Methods (ICATM). Considering all available evidence and input from regulators and stakeholders, the IL-8 Luc assay is considered useful as part of IATA to discriminate sensitisers from non-sensitisers for the purpose of hazard classification and labelling. Examples of the use of IL-8 Luc assay data in combination with other information are reported in the literature (4) (5) (6).

3. The IL-8 Luc assay proved to be transferable to laboratories experienced in cell culture and luciferase measurement. Within and between laboratory reproducibilities were 87.7% and 87.5%, respectively (2). Data generated in the validation study (2) and other published work (1) (6) show that versus the LLNA, the IL-8 Luc assay judged 118 out of 143 chemicals as positive or negative and judged 25 chemicals as inconclusive and the accuracy of the IL-8 Luc assay in distinguishing skin sensitisers (UN GHS/CLP Cat. 1) from non-sensitisers (UN GHS/CLP No Cat.) is 86% (101/118) with a sensitivity of 96% (92/96) and specificity of 41% (9/22). Excluding substances outside the applicability domain described below (paragraph 5), the IL-8 Luc assay judged 113 out of 136 chemicals as positive or negative and judged 23 chemicals as inconclusive and the accuracy of the IL-8 Luc assay is 89% (101/113) with sensitivity of 96% (92/96) and specificity of 53% (9/17). Using human data cited in Urbisch et al. (7), the IL-8 Luc assay judged 76 out of 90 chemicals as positive or negative and judged 14 chemicals as inconclusive and the accuracy is 80% (61/76), sensitivity is 93% (54/58) and specificity is 39% (7/18). Excluding substances outside the applicability domain, the IL-8 Luc assay judged 71 out of 84 chemicals as positive or negative and judged 13 chemicals as

inconclusive and the accuracy is 86% (61/71) with sensitivity of 93% (54/58) and specificity of 54% (7/13). False negative predictions with the IL-8 Luc assay are more likely to occur with chemicals showing low/moderate skin sensitisation potency (UN GHS/CLP subcategory 1B) than those with high potency (UN GHS/CLP subcategory 1A) (6). Together, the information supports a role for the IL-8 Luc assay in the identification of skin sensitisation hazards. The accuracy given for the IL-8 Luc assay as a standalone test is only for guidance, as the test should be considered in combination with other sources of information in the context of an IATA and in accordance with the provisions of paragraphs 7 and 8 in the General introduction. Furthermore, when evaluating non-animal tests for skin sensitisation, it should be remembered that the LLNA and other animal tests may not fully reflect the situation in humans.

4. On the basis of the data currently available, the IL-8 Luc assay was shown to be applicable to test chemicals covering a variety of organic functional groups, reaction mechanisms, skin sensitisation potency (as determined in *in vivo* studies) and physicochemical properties (2) (6).

5. Although the IL-8 Luc assay uses X-VIVO™ 15 as a solvent, it correctly evaluated chemicals with a Log K_{ow} >3.5 and those with a water solubility of around 100 µg/ ml as calculated by EPI Suite™ and its performance to detect sensitisers with poor water solubility is better than that of the IL-8 Luc assay using dimethyl sulfoxide (DMSO) as a solvent (2). However, negative results for test chemicals that are not dissolved at 20 mg/ml may produce false negative results due to their inability to dissolve in X-VIVO™ 15. Therefore, negative results for these chemicals should not be considered. A high false negative rate for anhydrides was seen in the validation study. Furthermore, because of the limited metabolic capability of the cell line (8) and the experimental conditions, pro-haptens (substances requiring metabolic activation) and pre-haptens (substances activated by air oxidation) might give negative results in the assay. However, although negative results for suspected pre/prohaptens should be interpreted with caution, the IL-8 Luc assay correctly judged 11 out of 11 pre-haptens, 6/6 pro-haptens, and 6/8 pre/pro-haptens in the IL-8 Luc assay data set (2). Based on the recent comprehensive review on three non-animal tests (the DPRA, the KeratinoSens™ and the h-CLAT) to detect pre and prohaptens (9), and based on the fact that THP-G8 cells used in the IL-8 Luc assay is a cell line derived from THP-1 that is used in the h-CLAT, the IL-8 Luc assay may also contribute to increase the sensitivity of non-animal tests to detect pre and pro-haptens in the combination of other tests. Surfactants tested so far gave (false) positive results irrespective of their type (e.g. cationic, anionic or on-ionic). Finally, chemicals that interfere with luciferase can confound its activity/measurement, causing apparent inhibition or increased luminescence (10). For example, phytoestrogen concentrations higher than 1µM were reported to interfere with luminescence signals in other luciferase-based reporter gene assays due to over-activation of the luciferase reporter gene.

Consequently, luciferase expression obtained at high concentrations of phytoestrogens or compounds suspected of producing phytoestrogen-like activation of the luciferase reporter gene needs to be examined carefully (11). Based on the above, surfactants, anhydrides and chemicals interfering with luciferase are outside the applicability domain of this assay. In cases where there is evidence demonstrating the non-applicability of the IL-8 Luc assay to other specific categories of test chemicals, the test should not be used for those specific categories.

6. As described above, the IL-8 Luc assay supports discrimination of skin sensitisers from non-sensitisers. Further work, preferably based on human data, is required to determine whether IL-8 Luc results can contribute to potency assessment when considered in combination with other information sources.

7. Definitions are provided in Appendix 3.1.

PRINCIPLE OF THE TEST

8. The IL-8 Luc assay makes use of a human monocytic leukemia cell line THP-1 that was obtained from the American Type Culture Collection (Manassas, VA, USA). Using this cell line, the Dept. of Dermatology, Tohoku University School of Medicine, established a THP-1-derived IL-8 reporter cell line, THP-G8, that harbours the Stable Luciferase Orange (SLO) and Stable Luciferase Red (SLR) luciferase genes under the control of the IL-8 and glyceraldehyde 3-phosphate dehydrogenase **(**GAPDH) promoters, respectively (1). This allows quantitative measurement of luciferase gene induction by detecting luminescence from well-established light producing luciferase substrates as an indicator of the activity of the IL-8 and GAPDH in cells following exposure to sensitising chemicals.

9. The dual-colour assay system comprises an orange-emitting luciferase (SLO; λmax = 580 nm) (12) for the gene expression of the IL-8 promoter as well as a red-emitting luciferase (SLR; λmax = 630 nm) (13) for the gene expression of the internal control promoter, GAPDH. The two luciferases emit different colours upon reacting with firefly D-luciferin and their luminescence is measured simultaneously in a one-step reaction by dividing the emission from the assay mixture using an optical filter (14) (Appendix 3.2).

10. THP-G8 cells are treated for 16 hours with the test chemical, after which SLO luciferase activity (SLO-LA) reflecting IL-8 promoter activity and SLR luciferase activity (SLR-LA) reflecting GAPDH promoter activity are measured. To make the abbreviations easy to understand, SLO-LA and SLR-LA are designated as IL8LA and GAPLA, respectively. Table 1 gives a description of the terms associated with luciferase activity in the IL-8 Luc assay. The measured values are used to calculate the normalised IL8LA (nIL8LA), which is the ratio of IL8LA to GAPLA; the induction of nIL8LA (Ind-IL8LA), which is the ratio of the arithmetic means of quadruple-measured values of the nIL8LA of THP-G8 cells

treated with a test chemical and the values of the nIL8LA of untreated THP-G8 cells; and the inhibition of GAPLA (Inh-GAPLA), which is the ratio of the arithmetic means of quadruple-measured values of the GAPLA of THP-G8 cells treated with a test chemical and the values of the GAPLA of untreated THP-G8 cells, and used as an indicator for cytotoxicity.

Table 1: Description of terms associated with the luciferase activity in the IL-8 Luc assay

Abbreviations	Definition
GAPLA	SLR luciferase activity reflecting GAPDH promoter activity
IL8LA	SLO luciferase activity reflecting IL-8 promoter activity
nIL8LA	IL8LA / GAPLA
Ind-IL8LA	nIL8LA of THP-G8 cells treated with chemicals / nIL8LA of untreated cells
Inh-GAPLA	GAPLA of THP-G8 treated with chemicals / GAPLA of untreated cells
CV05	The lowest concentration of the chemical at which Inh-GAPLA becomes < 0.05.

11. Performance standards (PS) (15) are available to facilitate the validation of modified *in vitro* IL-8 luciferase tests similar to the IL-8 Luc assay and allow for timely amendment of OECD Test Guideline 442E for their inclusion. OECD Mutual Acceptance of Data (MAD) will only be guaranteed for tests validated according to the PS, if these tests have been reviewed and included in Test Guideline 442E by the OECD (16).

DEMONSTRATION OF PROFICIENCY

12. Prior to routine use of the test described in this Appendix to test method B.71, laboratories should demonstrate technical proficiency, using the 10 Proficiency Substances listed in Appendix 3.3 in compliance with the Good *in vitro* Method Practices (17). Moreover, test users should maintain a historical database of data generated with the reactivity checks (see paragraph 15) and with the positive and solvent/vehicle controls (see paragraphs 21-24), and use these data to confirm the reproducibility of the test in their laboratory is maintained over time.

PROCEDURE

13. The Standard Operating Procedure (SOP) for the IL-8 Luc assay is available and should be employed when performing the test (18). Laboratories willing to perform the test can obtain the recombinant THP-G8 cell line from GPC Lab. Co. Ltd., Tottori, Japan, upon signing a Material Transfer Agreement (MTA) in line with the conditions of the OECD template. The following paragraphs provide a description of the main components and procedures of the assay.

Preparation of cells

14. The THP-G8 cell line from GPC Lab. Co. Ltd., Tottori, Japan, should be used for performing the IL-8 Luc assay (see paragraphs 8 and 13). On receipt, cells are propagated (2-4 passages) and stored frozen as a homogeneous stock. Cells from this stock can be propagated up to a maximum of 12 passages or a maximum of 6 weeks. The medium used for propagation is the RPMI-1640 culture medium containing 10% foetal bovine serum (FBS), antibiotic/antimycotic solution (100U/ml of penicillin G, 100µg/ml of streptomycin and 0.25µg/ml of amphotericin B in 0.85% saline) (e.g. GIBCO Cat#15240-062), 0.15µg/ml Puromycin (e.g. CAS:58-58-2) and 300µg/ml G418 (e.g. CAS:108321-42-2).

15. Prior to use for testing, the cells should be qualified by conducting a reactivity check. This check should be performed 1-2 weeks or 2-4 passages after thawing, using the positive control, 4-nitrobenzyl bromide (4-NBB) (CAS:100-11-8, ≥ 99% purity) and the negative control, lactic acid (LA) (CAS:50-21-5, ≥85% purity). 4-NBB should produce a positive response to Ind-IL8LA (≥1.4), while LA should produce a negative response to Ind-IL8LA (<1.4). Only cells that pass the reactivity check are used for the assay. The check should be performed according to the procedures described in paragraphs 22-24.

16. For testing, THP-G8 cells are seeded at a density of 2 to 5×10^5 cells/ml, and pre-cultured in culture flasks for 48 to 96 hours. On the day of the test, cells harvested from the culture flask are washed with RPMI-1640 containing 10% FBS without any antibiotics, and then, resuspended with RPMI-1640 containing 10% FBS without any antibiotics at 1×10^6 cells/ml. Then, cells are distributed into a 96-well flat-bottom black plate (e.g. Costar Cat#3603) with 50µl (5×10^4 cells/well).

Preparation of the test chemical and control substances

17. The test chemical and control substances are prepared on the day of testing. For the IL-8 Luc assay, test chemicals are dissolved in X-VIVO™ 15, a commercially available serum-free medium (Lonza, 04-418Q), to the final concentration of 20 mg/ml. X-VIVO™ 15 is added to 20 mg of test chemical (regardless of the chemical's solubility) in a microcentrifuge tube and brought to a volume of 1ml and then vortexed vigorously and shaken on a rotor at a maximum speed of 8 rpm for 30 min at an ambient temperature of about 20°C. Furthermore, if solid chemicals are still insoluble, the tube is sonicated until the chemical is dissolved completely or stably dispersed. For test chemicals soluble in X-VIVO™ 15, the solution is diluted by a factor of 5 with X-VIVO™ 15 and used as an X-VIVO™ 15 stock solution of the test chemical (4 mg/ml). For test chemicals not soluble in X-VIVO™ 15, the mixture is rotated again for at least 30 min, then centrifuged at 15,000 rpm (≈20 000*g*) for 5 min; the resulting supernatant is used as an X-VIVO™ 15 stock solution of the test chemical. A scientific rationale should be provided for the use of other solvents, such as DMSO, water, or the culture medium. The detailed procedure for dissolving chemicals is shown in Appendix 3.5. The X-VIVO™ 15 solutions described in

paragraphs 18-23 are mixed 1:1 (v/v) with the cell suspensions prepared in a 96-well flat-bottom black plate (see paragraph 16).

18. The first test run is aimed to determine the cytotoxic concentration and to examine the skin sensitising potential of chemicals. Using X-VIVO™ 15, serial dilutions of the X-VIVO™ 15 stock solutions of the test chemicals are made at a dilution factor of two (see Appendix 3.5) using a 96-well assay block (e.g. Costar Cat#EW-01729-03). Next, 50 µl/well of diluted solution is added to 50 µl of the cell suspension in a 96-well flat-bottom black plate. Thus for test chemicals that are soluble in X-VIVO ™ 15, the final concentrations of the test chemicals range from 0.002 to 2 mg/ml (Appendix 3.5). For test chemicals that are not soluble in X-VIVO ™ 15 at 20 mg/ml, only dilution factors that range from 2 to 2^{10}, are determined, although the actual final concentrations of the test chemicals remain uncertain and are dependent on the saturated concentration of the test chemicals in the X-VIVO ™ 15 stock solution.

19. In subsequent test runs (i.e. the second, third, and fourth replicates), the X-VIVO™ 15 stock solution is made at the concentration 4 times higher than the concentration of cell viability 05 (CV05; the lowest concentration at which the Inh-GAPLA becomes <0.05) in the first experiment. If Inh-GAPLA does not decrease below 0.05 at the highest concentration in the first run, the X-VIVO™ 15 stock solution is made at the first run highest concentration. The concentration of CV05 is calculated by dividing the concentration of the stock solution in the first run by dilution factor for CV05 (X) (dilution factor CV05 (X); the dilution factor required to dilute stock solution to CV05) (see Appendix 3.5). For test substances not soluble in X-VIVO at 20 mg/ml, CV05 is determined by the concentration of the stock solution x 1/X. For run 2 to 4, a second stock solution is prepared as 4 x CV50 (Appendix 3.5).

20. Serial dilutions of the X-VIVO™ 15 second stock solutions are made at a dilution factor of 1.5 using a 96-well assay block. Next, 50 µl/well of diluted solution is added to 50 µl of the cell suspension in the wells of a 96-well flat-bottom black plate. Each concentration of each test chemical should be tested in 4 wells. The samples are then mixed on a plate shaker and incubated for 16 hours at 37°C and 5% CO_2, after which the luciferase activity is measured as described below.

21. The solvent control is the mixture of 50 µl/well of X-VIVO™ 15 and 50 µl/well of cell suspension in RPMI-1640 containing 10% FBS.

22. The recommended positive control is 4-NBB. 20 mg of 4-NBB is prepared in a 1.5-ml microfuge tube, to which X-VIVO™ 15 is added up to 1 ml. The tube is vortexed vigorously and shaken on a rotor at a maximum speed of 8 rpm for at least 30 min. After centrifugation at 20 000g for 5 min, the supernatant is diluted by a factor of 4 with X-VIVO™ 15, and 500 µl of the diluted supernatant is transferred to a well in a 96-well assay block. The diluted supernatant is further diluted with X-VIVO™ 15 at factors of 2 and 4,

and 50 µl of the solution is added to 50 µl of THP-G8 cell suspension in the wells of a 96-well flat-bottom black plate (Appendix 3.6). Each concentration of the positive control should be tested in 4 wells. The plate is agitated on a plate shaker, and incubated in a CO_2 incubator for 16 hours (37°C, 5% CO_2), after which the luciferase activity is measured as described in paragraph 29.

23. The recommended negative control is LA. 20 mg of LA prepared in a 1.5-ml microfuge tube, to which X-VIVO™ 15 is added up to 1 ml (20 mg/ ml). Twenty mg/ml of LA solution is diluted by a factor of 5 with X-VIVO™ 15 (4 mg/ml); 500 µl of this 4 mg/ml LA solution is transferred to a well of a 96-well assay block. This solution is diluted by a factor of 2 with X-VIVO™ 15 and then diluted again by a factor of 2 to produce 2 mg/ml and 1 mg/ml solutions. 50 µl of these 3 solutions and vehicle control (X-VIVO™ 15) are added to 50 µl of THP-G8 cell suspension in the wells of a 96-well flat-bottom black plate. Each concentration of the negative control is tested in 4 wells. The plate is agitated on a plate shaker and incubated in a CO_2 incubator for 16 hours (37°C, 5% CO_2), after which the luciferase activity is measured as described in paragraph 29.

24. Other suitable positive or negative controls may be used if historical data are available to derive comparable run acceptance criteria.

25. Care should be taken to avoid evaporation of volatile test chemicals and cross-contamination between wells by test chemicals, e.g. by sealing the plate prior to the incubation with the test chemicals.

26. The test chemicals and solvent control require 2 to 4 runs to derive a positive or negative prediction (see Table 2). Each run is performed on a different day with fresh X-VIVO™ 15 stock solution of test chemicals and independently harvested cells. Cells may come from the same passage.

Luciferase activity measurements

27. Luminescence is measured using a 96-well microplate luminometer equipped with optical filters, e.g. Phelios (ATTO, Tokyo, Japan), Tristan 941 (Berthold, Bad Wildbad, Germany) and the ARVO series (PerkinElmer, Waltham, MA, USA). The luminometer must be calibrated for each test to ensure reproducibility (19). Recombinant orange and red emitting luciferases are available for this calibration.

28. 100µl of pre-warmed Tripluc® Luciferase assay reagent (Tripluc) is transferred to each well of the plate containing the cell suspension treated with or without chemical. The plate is shaken for 10 min at an ambient temperature of about 20°C. The plate is placed in the luminometer to measure the luciferase activity. Bioluminescence is measured for 3 sec each in the absence (F0) and presence (F1) of the optical filter. Justification should be provided for the use of alternative settings, e.g. depending on the model of luminometer used.

29. Parameters for each concentration are calculated from the measured values, e.g. IL8LA, GAPLA, nIL8LA, Ind-IL8LA, Inh-GAPLA, the mean ±SD of IL8LA, the mean ±SD of GAPLA, the mean ±SD of nIL8LA, the mean ±SD of Ind-IL8LA, the mean ±SD of Inh-GAPLA, and the 95% confidence interval of Ind-IL8LA. Definitions of the parameters used in this paragraph are provided in Appendices I and IV, respectively.

30. Prior to measurement, colour discrimination in multi-colour reporter assays is generally achieved using detectors (luminometer and plate reader) equipped with optical filters, such as sharp-cut (long-pass or short-pass) filters or band-pass filters. The transmission coefficients of the filters for each bioluminescence signal colour should be calibrated prior to testing, per Appendix 3.2.

DATA AND REPORTING

Data evaluation

31. Criteria for a positive/negative decision require that in each run:

- an IL-8 Luc assay prediction is judged positive if a test chemical has a Ind-IL8LA ≥ 1.4 and the lower limit of the 95% confidence interval of Ind-IL8LA ≥ 1.0
- an IL-8 Luc assay prediction is judged negative if a test chemical has a Ind-IL8LA < 1.4 and/or the lower limit of the 95% confidence interval of Ind-IL8LA < 1.0

Prediction model

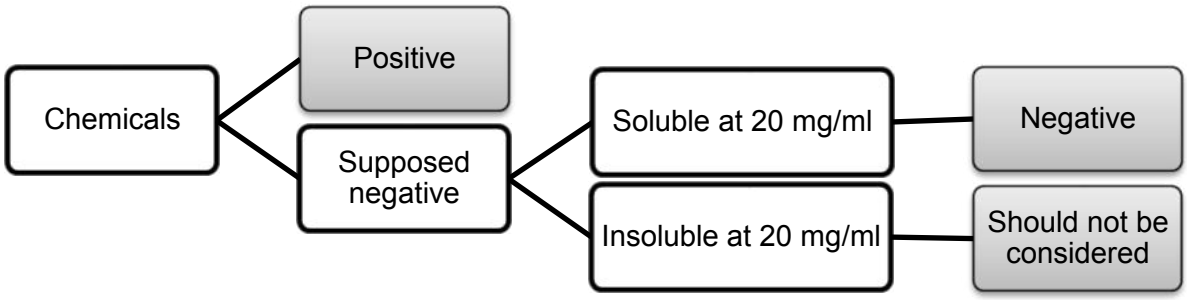
32. Test chemicals that provide two positive results from among the 1st, 2nd, 3rd or 4th runs are identified as positives whereas those that give three negative results from among the 1st, 2nd, 3rd or 4th runs are identified as supposed negative (Table 2). Among supposed negative chemicals, chemicals that are dissolved at 20 mg/ml of X-VOVO™ 15 are judged as negative, while chemicals that are not dissolved at 20 mg/ml of X-VOVO™ 15 should not be considered (Figure 1).

Table 2: Criteria for identifying positive and supposed negative

1st run	2nd run	3rd run	4th run	Final prediction
Positive	Positive	-	-	Positive
	Negative	Positive	-	Positive
		Negative	Positive	Positive
			Negative	Supposed negative
Negative	Positive	Positive	-	Positive
		Negative	Positive	Positive
			Negative	Supposed negative
	Negative	Positive	Positive	Positive
			Negative	Supposed negative

	Negative	-	Supposed negative

Figure 1: Prediction model for final judgment



Acceptance criteria

33. The following acceptance criteria should be met when using the IL-8 Luc assay:

- Ind-IL8LA should be more than 5.0 at least in one concentration of the positive control, 4-NBB, in each run.
- Ind-IL8LA should be less than 1.4 at any concentration of the negative control, lactic acid, in each run.
- Data from plates for which the GAPLA of control wells with cells and Tripluc but without chemicals is less than 5 times of that of well containing test medium only (50 µl/well of RPMI-1640 containing 10% FBS and 50 µl/well of X-VIVO^{TM} 15) should be rejected.
- Data from plates for which the Inh-GAPLA of all concentrations of the test or control chemicals is less than 0.05 should be rejected. In this case, the first test should be repeated so the highest final concentration of the repeated test is the lowest final concentration of the previous test.

Test report

34. The test report should include the following information:

Test chemicals

Mono-constituent substance:

- Chemical identification, such as IUPAC or CAS name(s), CAS number(s), SMILES or InChI code, structural formula, and/or other identifiers;
- Physical appearance, water solubility, molecular weight, and additional relevant physicochemical properties, to the extent available;
- Purity, chemical identity of impurities as appropriate and practically feasible, etc.;
- Treatment prior to testing, if applicable (e.g. warming, grinding);
- Solubility in X-VIVO™ 15. For chemicals that are insoluble in X-VIVO™ 15, whether precipitation or flotation are observed after centrifugation;
- Concentration(s) tested;
- Storage conditions and stability to the extent available;
- Justification for choice of solvent/vehicle for each test chemical if X-VIVO™ 15 has not been used.

Multi-constituent substance, UVCB and mixture:

- Characterisation as far as possible by e.g. chemical identity (see above), purity, quantitative occurrence and relevant physicochemical properties (see above) of the constituents, to the extent available;
- Physical appearance, water solubility, and additional relevant physicochemical properties, to the extent available;
- Molecular weight or apparent molecular weight in case of mixtures/polymers of known compositions or other information relevant for the conduct of the study;
- Treatment prior to testing, if applicable (e.g. warming, grinding);
- Solubility in X-VIVO™ 15. For chemicals that are insoluble in X-VIVO™ 15, whether precipitation or flotation are observed after centrifugation;
- Concentration(s) tested;
- Storage conditions and stability to the extent available.
- Justification for choice of solvent/vehicle for each test chemical, if X-VIVO™ 15 has not been used.

Controls

Positive control:

- Chemical identification, such as IUPAC or CAS name(s), CAS number(s), SMILES or InChI code, structural formula, and/or other identifiers;

- Physical appearance, water solubility, molecular weight, and additional relevant physicochemical properties, to the extent available and where applicable;
- Purity, chemical identity of impurities as appropriate and practically feasible, etc;
- Treatment prior to testing, if applicable (e.g. warming, grinding);
- Concentration(s) tested;
- Storage conditions and stability to the extent available;
- Reference to historical positive control results demonstrating suitable acceptance criteria, if applicable.

Negative control:

- Chemical identification, such as IUPAC or CAS name(s), CAS number(s), and/or other identifiers;
- Purity, chemical identity of impurities as appropriate and practically feasible, etc;
- Physical appearance, molecular weight, and additional relevant physicochemical properties in the case other negative controls than those mentioned in the Test Guideline are used and to the extent available;
- Storage conditions and stability to the extent available;
- Justification for choice of solvent for each test chemical.

Test conditions

- Name and address of the sponsor, test facility and study director;
- Description of test used;
- Cell line used, its storage conditions, and source (e.g. the facility from which it was obtained);
- Lot number and origin of FBC, supplier name, lot number of 96-well flat-bottom black plate, and lot number of Tripluc reagent;
- Passage number and cell density used for testing;
- Cell counting method used for seeding prior to testing and measures taken to ensure homogeneous cell number distribution;
- Luminometer used (e.g. model), including instrument settings, luciferase substrate used, and demonstration of appropriate luminescence measurements based on the control test described in Appendix 3.2;

- The procedure used to demonstrate proficiency of the laboratory in performing the test (e.g. by testing of proficiency substances) or to demonstrate reproducible performance of the test over time.

Test procedure

- Number of replicates and runs performed;
- Test chemical concentrations, application procedure and exposure time (if different from those recommended);
- Description of evaluation and decision criteria used;
- Description of study acceptance criteria used;
- Description of any modifications of the test procedure.

Results

- Measurements of IL8LA and GAPLA;
- Calculations for nIL8LA, Ind-IL8LA, and Inh-GAPLA;
- The 95% confidence interval of Ind-IL8LA;
- A graph depicting dose-response curves for induction of luciferase activity and viability;
- Description of any other relevant observations, if applicable.

Discussion of the results

- Discussion of the results obtained with the IL-8 Luc assay;
- Consideration of the assay results in the context of an IATA, if other relevant information is available.

Conclusion

LITERATURE

(1) Takahashi T, Kimura Y, Saito R, Nakajima Y, Ohmiya Y, Yamasaki K, and Aiba S. (2011). An *in vitro* test to screen skin sensitizers using a stable THP-1-derived IL-8 reporter cell line, THP-G8. Toxicol Sci 124:359-69.

(2) OECD (2017). Validation report for the international validation study on the IL-8 Luc assay as a test evaluating the skin sensitizing potential of chemicals conducted by the IL-8 Luc Assay. Series on Testing and Assessment No 267, ENV/JM/MONO(2017)19. Organisation for Economic Cooperation and Development, Paris. Available at: http://www.oecd.org/env/ehs/testing/series-testing-assessment-publications-number.htm.

(3) OECD (2017). Report of the Peer Review Panel for the IL-8 Luciferase (IL-8 Luc) Assay for *in vitro* skin sensitisation. Series on Testing and Assessment No 258, ENV/JM/MONO(2017)20. Organisation for Economic Cooperation and Development, Paris. Available at: http://www.oecd.org/env/ehs/testing/series-testing-assessment-publications-number.htm.

(4) OECD (2016) Guidance Document On The Reporting Of Defined Approaches And Individual Information Sources To Be Used Within Integrated Approaches To Testing And Assessment (IATA) For Skin Sensitisation, Series on Testing & Assessment No 256, ENV/JM/MONO(2016)29. Organisation for Economic Cooperation and Development, Paris. Available at: http://www.oecd.org/env/ehs/testing/series-testing-assessment-publications-number.htm.

(5) van der Veen JW, Rorije E, Emter R, Natsch A, van Loveren H, and Ezendam J. (2014). Evaluating the performance of integrated approaches for hazard identification of skin sensitizing chemicals. Regul Toxicol Pharmacol 69:371-9.

(6) Kimura Y, Fujimura C, Ito Y, Takahashi T, Nakajima Y, Ohmiya Y, and Aiba S. (2015). Optimization of the IL-8 Luc assay as an *in vitro* test for skin sensitization. Toxicol *In Vitro* 29:1816-30.

(7) Urbisch D, Mehling A, Guth K, Ramirez T, Honarvar N, Kolle S, Landsiedel R, Jaworska J, Kern PS, Gerberick F, et al. (2015). Assessing skin sensitization hazard in mice and men using non-animal test methods. Regul Toxicol Pharmacol 71:337-51.

(8) Ashikaga T, Sakaguchi H, Sono S, Kosaka N, Ishikawa M, Nukada Y, Miyazawa M, Ito Y, Nishiyama N, and Itagaki H. (2010). A comparative evaluation of *in vitro* skin sensitisation tests: the human cell-line activation

test (h-CLAT) versus the local lymph node assay (LLNA). Alternatives to laboratory animals: ATLA 38:275-84.

(9) Patlewicz G, Casati S, Basketter DA, Asturiol D, Roberts DW, Lepoittevin J-P, Worth A and Aschberger K (2016) Can currently available non-animal methods detect pre and pro haptens relevant for skin sensitisation? Regul Toxicol Pharmacol, 82:147-155.

(10) Thorne N, Inglese J, and Auld DS. (2010). Illuminating insights into firefly luciferase and other bioluminescent reporters used in chemical biology. Chem Biol 17:646-57.

(11) OECD (2016). Test No 455: Performance-Based Test Guideline for Stably Transfected Transactivation *In Vitro* Assays to Detect Estrogen Receptor Agonists and Antagonists, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264265295-en.

(12) Viviani V, Uchida A, Suenaga N, Ryufuku M, and Ohmiya Y. (2001). Thr226 is a key residue for bioluminescence spectra determination in beetle luciferases. Biochem Biophys Res Commun 280:1286-91.

(13) Viviani VR, Bechara EJ, and Ohmiya Y. (1999). Cloning, sequence analysis, and expression of active Phrixothrix railroad-worms luciferases: relationship between bioluminescence spectra and primary structures. Biochemistry 38:8271-9.

(14) Nakajima Y, Kimura T, Sugata K, Enomoto T, Asakawa A, Kubota H, Ikeda M, and Ohmiya Y. (2005). Multicolor luciferase assay system: one-step monitoring of multiple gene expressions with a single substrate. Biotechniques 38:891-4.

(15) OECD (2017). To be published - Performance Standards for the assessment of proposed similar or modified *in vitro* skin sensitisation IL-8 luc test methods. OECD Environment, Health and Safety Publications, Series on Testing and Assessment. OECD, Paris, France

(16) OECD (2005). Guidance Document the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment. OECD Environment, Health and Safety publications, OECD Series on Testing and Assessment No 34. OECD, Paris, France.

(17) OECD (2018). Draft Guidance document: Good *In Vitro* Method Practices (GIVIMP) for the Development and Implementation of *In Vitro* Methods for Regulatory Use in Human Safety Assessment. Organisation for Economic Cooperation and Development, Paris. Available at: http://www.oecd.org/env/ehs/testing/OECD Final Draft GIVIMP.pdf.

(18) JaCVAM (2016). IL-8 Luc assay protocol, Available at. http://www.jacvam.jp/en_effort/effort02.html.

(19) Niwa K, Ichino Y, Kumata S, Nakajima Y, Hiraishi Y, Kato D, Viviani VR, and Ohmiya Y. (2010). Quantum yields and kinetics of the firefly bioluminescence reaction of beetle luciferases. Photochem Photobiol 86:1046-9.

(20) OECD (2012). The Adverse Outcome Pathway for Skin Sensitisation Initiated by Covalent Binding to Proteins, Part 1: Scientific Evidence. OECD Environment, Health and Safety Publications, Series on Testing and Assessment No 168. OECD, Paris, France.

(21) United Nations (2015). Globally Harmonized System of Classification and Labelling of Chemicals (GHS). Sixth revised edition. New York & Geneva: United Nations Publications. ISBN: 978-92-1-117006-1. Available at: http://www.unece.org/trans/danger/publi/ghs/ghs_rev05/05files_e.html.

Appendix 3.1

DEFINITIONS

Accuracy: The closeness of agreement between test results and accepted reference values. It is a measure of test performance and one aspect of relevance. The term is often used interchangeably with concordance to mean the proportion of correct outcomes of a test (16).

AOP (Adverse Outcome Pathway): Sequence of events from the chemical structure of a target chemical or group of similar chemicals through the molecular initiating event to an *in vivo* outcome of interest (20).

Chemical: A substance or a mixture.

CV05: Cell viability 05, i.e. minimum concentration at which chemicals show less than 0.05 of Inh-GAPLA.

FInSLO-LA: Abbreviation used in the validation report and in previous publications regarding the IL-8 Luc assay to refer to Ind-IL8LA. See Ind-IL8LA for definition.

GAPLA: Luciferase Activity of Stable Luciferase Red (SLR) (λmax = 630 nm), regulated by GAPDH promoter and demonstrates cell viability and viable cell number.

Hazard: Inherent property of an agent or situation having the potential to cause adverse effects when an organism, system or (sub) population is exposed to that agent.

IATA (Integrated Approach to Testing and Assessment): A structured approach used for hazard identification (potential), hazard characterisation (potency) and/or safety assessment (potential/potency and exposure) of a chemical or group of chemicals, which strategically integrates and weights all relevant data to inform regulatory decision regarding potential hazard and/or risk and/or the need for further targeted and therefore minimal testing.

II-SLR-LA: Abbreviation used in the validation report and in previous publications regarding the IL-8 Luc assay to refer to Inh-GAPLA. See Inh-GAPLA for definition

IL-8 (Interleukin-8): A cytokine derived from endothelial cells, fibroblasts, keratinocytes, macrophages, and monocytes that causes chemotaxis of neutrophils and T-cell lymphocytes.

IL8LA: Luciferase Activity of Stable Luciferase Orange (SLO) (λmax = 580 nm), regulated by IL-8 promoter.

Ind-IL8LA: Fold induction of IL8LA. It is obtained by dividing the nIL8LA of THP-G8 cells treated with chemicals by that of non-stimulated THP-G8 cells and represents the induction of IL-8 promoter activity by chemicals.

Inh-GAPLA: Inhibition of GAPLA. It is obtained by dividing GAPLA of THP-G8 treated with chemicals with GAPLA of non-treated THP-G8 and represents cytotoxicity of chemicals.

Minimum induction threshold (MIT): the lowest concentration at which a chemical satisfies the positive criteria

Mixture: A mixture or a solution composed of two or more substances.

Mono-constituent substance: A substance, defined by its quantitative composition, in which one main constituent is present to at least 80% (w/w).

Multi-constituent substance: A substance, defined by its quantitative composition, in which more than one of the main constituents is present in a concentration ≥ 10% (w/w) and < 80% (w/w). A multi-constituent substance is the result of a manufacturing process. The difference between mixture and multi-constituent substance is that a mixture is obtained by blending of two or more substances without chemical reaction. A multi-constituent substance is the result of a chemical reaction.

nIL8LA: The SLO luciferase activity reflecting IL-8 promoter activity (IL8LA) normalised by the SLR luciferase activity reflecting GAPDH promoter activity (GALPA). It represents IL-8 promoter activity after considering cell viability or cell number.

nSLO-LA: Abbreviation used in the validation report and in previous publications regarding the IL-8 Luc assay to refer to nIL8LA. See nIL8LA for definition

Positive control: A replicate containing all components of a test system and treated with a substance known to induce a positive response. To ensure that variability in the positive control response across time can be assessed, the magnitude of the positive response should not be excessive.

Pre-haptens: Chemicals which become sensitisers through abiotic transformation.

Pro-haptens: Chemicals requiring enzymatic activation to exert skin sensitisation potential.

Relevance: Description of relationship of the test to the effect of interest and whether it is meaningful and useful for a particular purpose. It is the extent to which the test correctly measures or predicts the biological effect of interest. Relevance incorporates consideration of the accuracy (concordance) of a test (16).

Reliability: Measures of the extent that a test can be performed reproducibly within and between laboratories over time, when performed using the same protocol. It is assessed by calculating intra- and inter-laboratory reproducibility and intra-laboratory repeatability (16).

Run: A run consists of one or more test chemicals tested concurrently with a solvent/vehicle control and with a positive control.

Sensitivity: The proportion of all positive/active chemicals that are correctly classified by the test. It is a measure of accuracy for a test that produces categorical results, and is an important consideration in assessing the relevance of a test (16).

SLO-LA: Abbreviation used in the validation report and in previous publications regarding

the IL-8 Luc assay to refer to IL8LA. See IL8LA for definition.

SLR-LA: Abbreviation used in the validation report and in previous publications regarding the IL-8 Luc assay to refer to GAPLA. See GAPLA for definition.

Solvent/vehicle control: An untreated sample containing all components of a test system except of the test chemical, but including the solvent/vehicle that is used. It is used to establish the baseline response for the samples treated with the test chemical dissolved or stably dispersed in the same solvent/vehicle. When tested with a concurrent medium control, this sample also demonstrates whether the solvent/vehicle interacts with the test system.

Specificity: The proportion of all negative/inactive chemicals that are correctly classified by the test. It is a measure of accuracy for a test that produces categorical results and is an important consideration in assessing the relevance of a test (16).

Substance: A chemical elements and its compounds in the natural state or obtained by any production manufacturing process, including any additive necessary to preserve the its stability of the product and any impurities deriving from the process used, but excluding any solvent which may be separated without affecting the stability of the substance or changing its composition.

Surfactant: Also called surface-active agent, this is a substance, such as a detergent, that can reduce the surface tension of a liquid and thus allow it to foam or penetrate solids; it is also known as a wetting agent. (TG437)

Test chemical: Any substance or mixture tested using this method.

THP-G8: An IL-8 reporter cell line used in IL-8 Luc assay. The human macrophage-like cell line THP-1 was transfected the SLO and SLR luciferase genes under the control of the IL-8 and GAPDH promoters, respectively.

United Nations Globally Harmonized System of Classification and Labeling of Chemicals (UN GHS): A system proposing the classification of chemicals (substances and mixtures) according to standardised types and levels of physical, health and environmental hazards, and addressing corresponding communication elements, such as pictograms, signal words, hazard statements, precautionary statements and safety data sheets, so that to convey information on their adverse effects with a view to protect people (including employers, workers, transporters, consumers and emergency responders) and the environment (21).

UVCB: substances of unknown or variable composition, complex reaction products or biological materials.

Valid test method: A test considered to have sufficient relevance and reliability for a specific purpose and which is based on scientifically sound principles. A test is never valid in an absolute sense, but only in relation to a defined purpose.

Appendix 3.2

PRINCIPLE OF MEASUREMENT OF LUCIFERASE ACTIVITY AND DETERMINATION OF THE TRANSMISSION COEFFICIENTS OF OPTICAL FILTER FOR SLO AND SLR

MultiReporter Assay System -Tripluc- can be used with a microplate-type luminometer with a multi-colour detection system, which can equip an optical filter (e.g. Phelios AB-2350 (ATTO), ARVO (PerkinElmer), Tristar LB941 (Berthold)). The optical filter used in measurement is 600–620 nm long or short pass filter, or 600–700 nm band pass filter.

Measurement of two-colour luciferases with an optical filter.

This is an example using Phelios AB-2350 (ATTO). This luminometer is equipped with a 600 nm long pass filter (R60 HOYA Co., 600 nm LP, Filter 1) for splitting SLO (λmax = 580 nm) and SLR (λmax = 630 nm) luminescence.

To determine transmission coefficients of the 600 nm LP, first, using purified SLO and SLR luciferase enzymes, measure i) the SLO and SLR bioluminescence intensity without filter (F0), ii) the SLO and SLR bioluminescence intensity that passed through 600 nm LP (Filter 1), and iii) calculate the transmission coefficients of 600 nm LP for SLO and SLR listed below.

Transmission coefficients		Abbreviation	Definition
SLO	Filter 1 Transmission coefficients	κO_{R60}	The filter's transmission coefficient for the SLO
SLR	Filter 1 Transmission coefficients	κR_{R60}	The filter's transmission coefficient for the SLR

When the intensity of SLO and SLR in test sample are defined as O and R, respectively, i) the intensity of light without filter (all optical) F0 and ii) the intensity of light that transmits through 600 nm LP (Filter 1) F1 are described as below.

$$F0=O+R$$

$$F1=\kappa O_{R60} \times O + \kappa R_{R60} \times R$$

These formulas can be rephrased as follows:

$$\begin{pmatrix} F0 \\ F1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ \kappa O_{R60} & \kappa R_{R60} \end{pmatrix} \begin{pmatrix} O \\ R \end{pmatrix}$$

Then using calculated transmittance factors (κO_{R60} and κR_{R60}) and measured F0 and F1, you can calculate O and R-value as follows:

$$\begin{pmatrix} O \\ R \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ \kappa O_{R60} & \kappa R_{R60} \end{pmatrix}^{-1} \begin{pmatrix} F0 \\ F1 \end{pmatrix}$$

Materials and methods for determining transmittance factor

(1) Reagents

Single purified luciferase enzymes:

Lyophilised purified SLO enzyme

Lyophilised purified SLR enzyme

(which for the validation work were obtained from GPC Lab. Co. Ltd., Tottori, Japan with THP-G8 cell line)

Assay reagent:

Tripluc® Luciferase assay reagent (for example from TOYOBO Cat#MRA-301)

Medium: for luciferase assay (30 ml, stored at 2 – 8°C)

Reagent	Conc.	Final conc. in medium	Required amount
RPMI-1640	-	-	27 ml
FBS	-	10 %	3 ml

(2) Preparation of enzyme solution

Dissolve lyophilised purified luciferase enzyme in tube by adding 200 µl of 10 ~ 100 mM Tris/HCl or Hepes/HCl (pH 7.5 ~ 8.0) supplemented with 10% (w/v) glycerol, divide the enzyme solution into 10 µl aliquots in 1.5 ml disposable tubes and store them in a freezer at -80°C. The frozen enzyme solution can be used for up to 6 months. When used, add 1 ml of medium for luciferase assay (RPMI-1640 with 10% FBS) to each tube containing the enzyme solutions (diluted enzyme solution) and keep them on ice to prevent deactivation.

(3) Bioluminescence measurement

Thaw Tripluc® Luciferase assay reagent (Tripluc) and keep it at room temperature either in a water bath or at ambient air temperature. Power on the luminometer 30 min before starting the measurement to allow the photomultiplier to stabilise. Transfer 100 µl of the diluted enzyme solution to a black 96 well plate (flat bottom) (the SLO reference sample to #B1, #B2, #B3, the SLR reference sample to #D1, #D2, #D3). Then, transfer 100 µl of pre-warmed Tripluc to each well of the plate containing the diluted enzyme solution using a pipetman. Shake the plate for 10 min at room temperature (about 25°C) using a plate shaker. Remove bubbles from the solutions in wells if they appear. Place the plate in the luminometer to measure the luciferase activity. Bioluminescence is measured for 3 sec each in the absence (F0) and presence

(F1) of the optical filter.

Transmission coefficient of the optical filter was calculated as follows:

Transmission coefficient (SLO (κO_{R60}))= (#B1 of F1+ #B2 of F1+ #B3 of F1) / (#B1 of F0+ #B2 of F0+ #B3 of F0)

Transmission coefficient (SLR (κR_{R60}))= (#D1 of F1+ #D2 of F1+ #D3 of F1) / (#D1 of F0+ #D2 of F0+ #D3 of F0)

Calculated transmittance factors are used for all the measurements executed using the same luminometer.

Quality control of equipment

The procedures described in the IL-8 Luc protocol should be used (18).

Appendix 3.3

PROFICIENCY SUBSTANCES

Prior to routine use of the test described in this Appendix to test method B.71, laboratories should demonstrate technical proficiency by obtaining the expected IL-8 Luc assay prediction for the 10 substances recommended in Table 1 and by obtaining values that fall within the respective reference range for at least 8 out of the 10 proficiency substances (selected to represent the range of responses for skin sensitisation hazards). Other selection criteria were that the substances are commercially available, and that high-quality *in vivo* reference data as well as high quality *in vitro* data generated with the IL-8 Luc assay are available. Also, published reference data are available for the IL-8 Luc assay (6) (1).

Table 1: Recommended substances for demonstrating technical proficiency with the IL-8 Luc assay

Proficiency substances	**CAS no.**	**State**	**Solubility in X-VIVO15 at 20 mg/ml**	***In vivo* prediction[1]**	**IL-8 Luc prediction[2]**	**Reference range (µg/ml) [3]**	
						CV05[4]	**IL-8 Luc MIT[5]**
2,4-Dinitrochlorobenzene	97-00-7	Solid	Insoluble	Sensitiser (Extreme)	Positive	2.3-3.9	0.5-2.3
Formaldehyde	50-00-0	Liquid	Soluble	Sensitiser (Strong)	Positive	9-30	4-9
2-Mercaptobenzothiazole	149-30-4	Solid	Insoluble	Sensitiser (Moderate)	Positive	250-290	60-250
Ethylenediamine	107-15-3	Liquid	Soluble	Sensitiser (Moderate)	Positive	500-700	0.1-0.4
Ethyleneglycol dimethacrylate	97-90-5	Liquid	Insoluble	Sensitiser (Weak)	Positive	>2000	0.04-0.1
4-Allylanisole (Estragol)	140-67-0	Liquid	Insoluble	Sensitiser (Weak)	Positive	>2000	0.01-0.07
Streptomycin sulphate	3810-74-0	Solid	Soluble	Non-sensitiser	Negative	>2000	>2000
Glycerol	56-81-5	Liquid	Soluble	Non-sensitiser	Negative	>2000	>2000
Isopropanol	67-63-0	Liquid	Soluble	Non-sensitiser	Negative	>2000	>2000

Abbreviations: CAS no. = Chemical Abstracts Service Registry Number

[1] The *in vivo* potency is derived using the criteria proposed by ECETOC (19).

[2] Based on historical observed values (1) (6).

[3] CV05 and IL-8 Luc MIT were calculated using water solubility given by EPI SuiteTM.

[4] CV05: the minimum concentration at which chemicals show less than 0.05 of Inh-GAPLA.

[5] MIT: the lowest concentrations at which a chemical satisfies the positive criteria.

Appendix 3.4

INDEXES AND JUDGMENT CRITERIA

nIL8LA (nSLO-LA)

The j-th repetition (j = 1-4) of the i-th concentration (i = 0-11) is measured for IL8LA (SLO-LA) and GAPLA (SLR-LA) respectively. The normalised IL8LA, referred to as nIL8LA (nSLO-LA), and is defined as:

$$\mathrm{nIL8LA_{ij}} = \mathrm{IL8LA_{ij}}/\mathrm{GAPLA_{ij}}$$

This is the basic unit of measurement in this assay.

Ind-IL8LA (FInSLO-LA)

The fold increase of the averaged nIL8LA (nSLO-LA) for the repetition on the i-th concentration compared with it at the 0 concentration, Ind-IL8LA, is the primary measure of this assay. This ratio is written by the following formula:

$$\mathrm{Ind-IL8LA_i} = \{(1/4) \times \textstyle\sum_j \mathrm{nIL8LA_{ij}}\}/\{(1/4) \times \textstyle\sum_j \mathrm{nIL8LA_{0j}}\}$$

The lead laboratory has proposed that a value of 1.4 corresponds to a positive result for the tested chemical. This value is based on the investigation of the historical data of the lead laboratory. Data management team then used this value through all the phases of validation study. The primary outcome, Ind-IL8LA, is the ratio of 2 arithmetic means as shown in equation.

95% confidence interval (95% CI)

The 95% confidence interval (95% CI) based on the ratio can be estimated to show the precision of this primary outcome measure. The lower limit of the 95% CI ≥ 1 indicates that the nIL8LA with the i-th concentration is significantly greater than that with solvent control. There are several ways to construct the 95% CI. We used the method known as Fieller's theorem in this study. This 95% confidence interval theorem is obtained from the following formula:

$$\left[\frac{-\mathrm{B} - \sqrt{\mathrm{B^2} - 4\mathrm{AC}}}{2\mathrm{A}}, \frac{-\mathrm{B} + \sqrt{\mathrm{B^2} - 4\mathrm{AC}}}{2\mathrm{A}}\right],$$

Where

$$\mathrm{A} = \bar{\mathrm{x}}_0^2 - \mathrm{t}_{0.975(\nu)}^2 \times \frac{\mathrm{sd}_0^2}{\mathrm{n}_0}, \qquad \mathrm{B} = -2 \times \bar{\mathrm{x}} \times \bar{\mathrm{y}}, \qquad \mathrm{C} = \bar{\mathrm{y}}_\mathrm{i}^2 - \mathrm{t}_{0.975(\nu)}^2 \times \frac{\mathrm{sd}_{\mathrm{yi}}^2}{\mathrm{n}_{\mathrm{yi}}}, \text{and } \mathrm{n}_0 = 4,$$

$$\bar{\mathrm{x}}_0 = (1/\mathrm{n}_0) \times \textstyle\sum_\mathrm{j} \mathrm{nIL8LA_{0j}}, \quad \mathrm{sd}_0^2 = \{1/(\mathrm{n}_0 - 1)\} \times \textstyle\sum_\mathrm{j}\left(\mathrm{nIL8LA_{0j}} - \bar{\mathrm{x}}_0\right)^2,$$

$n_{yi} = 4,\ \bar{y}_i = (1/n_{yi}) \times \sum_j (nIL8LA_{ij}),\ sd_{yi}^2 = \{1/(n_{yj} - 1)\} \times \sum_j (nIL8LA_{ij} - \bar{y}_i)^2$.

$t_{0.975(\nu)}$ is 97.5 percentile of the central t distribution with the ν of the degree of freedom, where

$$\nu = \left(\frac{sd_0^2}{n_0} + \frac{sd_{yi}^2}{n_{yi}}\right) / \left\{\left(\frac{sd_0^2}{n_0}\right)^2 / (n_0 - 1) + \left(\frac{sd_{yi}^2}{n_{yi}}\right) / (n_{yi} - 1)\right\}.$$

Inh-GAPLA (II-SLR-LA)

The Inh-GAPLA is a ratio of the averaged GAPLA (SLR-LA) for the repetition of the i-th concentration compared with that with solvent control, and this is written by

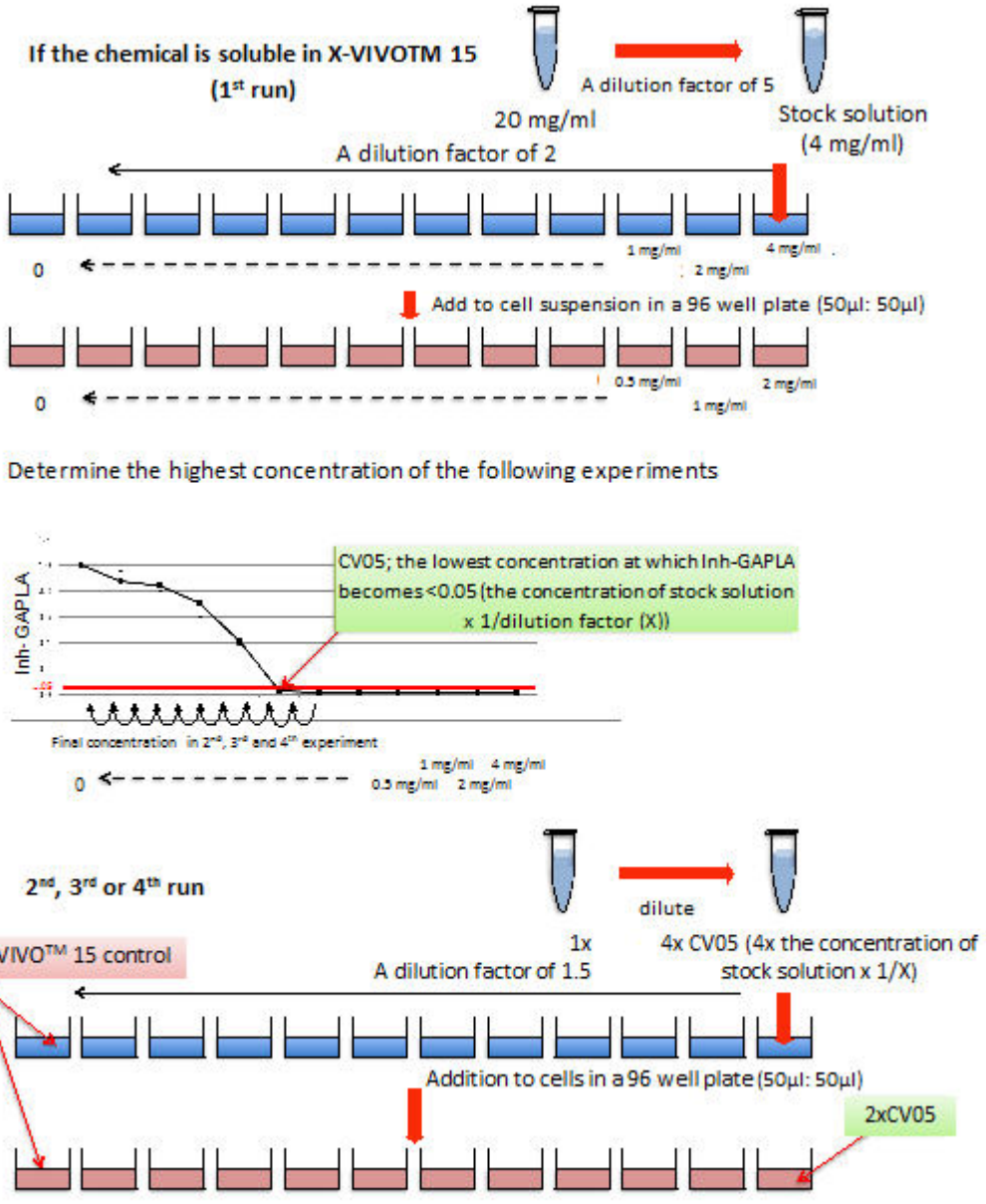
$$Inh - GAPLA_i = \{(1/4) \times \sum_j GAPLA_{ij}\} / \{(1/4) \times \sum_j GAPLA_{0j}\}.$$

Since the GAPLA is the denominator of the nIL8LA, an extremely small value causes large variation in the nIL8LA. Therefore, Ind-IL8LA values with an extremely small value of Inh-GAPLA (less than 0.05) might be considered poor precision.

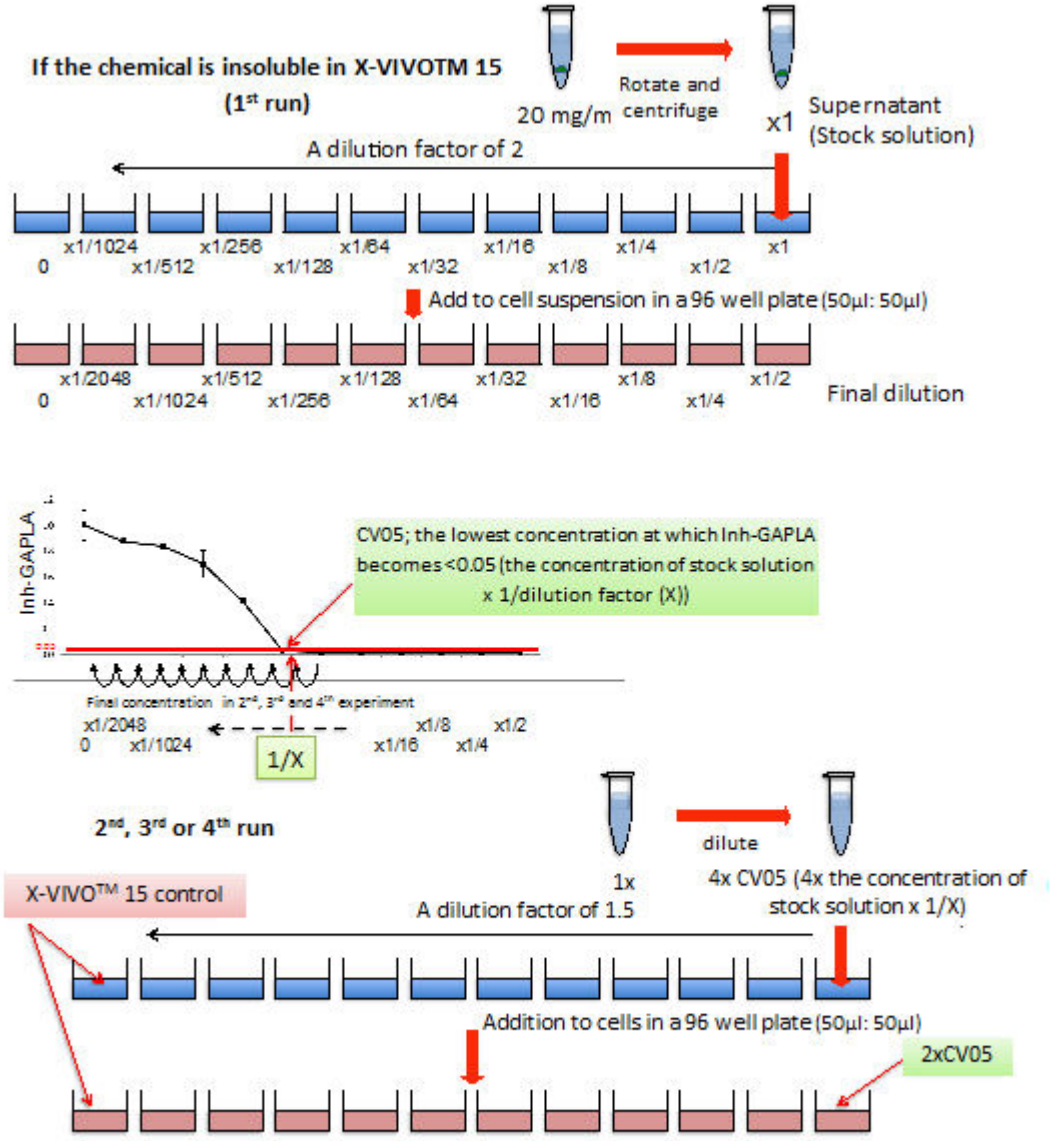
Appendix 3.5

THE SCHEME OF THE METHODS TO DISSOLVE CHEMICALS FOR THE IL-8 LUC ASSAY.

(a) For chemicals dissolved in X-VIVO™ 15 at 20 mg/ml

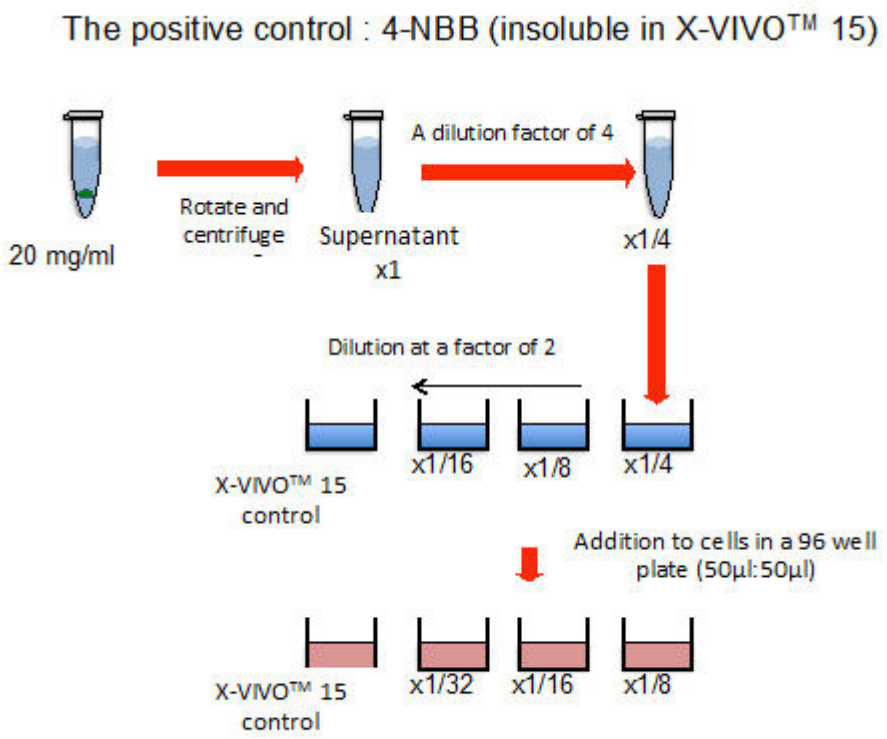


(b) For chemicals insoluble in X-VIVOTM 15 at 20 mg/ml



Appendix 3.6

THE SCHEME OF THE METHOD TO DISSOLVE 4-NBB FOR THE POSITIVE CONTROL OF THE IL-8 LUC ASSAY.



"

(9) In Part C, the following Chapters are added:

"C.52 MEDAKA EXTENDED ONE GENERATION REPRODUCTION TEST (MEOGRT)

INTRODUCTION

1. This test method is equivalent to OECD test guideline (TG) 240 (2015). The Medaka Extended One Generation Test (MEOGRT) describes a comprehensive test method based on fish exposed over multiple generations to give data relevant to ecological hazard and risk assessment of chemicals, including suspected endocrine disrupting chemicals (EDCs). Exposure in the MEOGRT continues until hatching (until two weeks post fertilisation, wpf) in the second (F2) generation. Additional investigations would be needed to justify the utility of extending the F2 generation beyond hatching; at this time, there is insufficient information to provide relevant conditions or criteria for warranting the extension of the F2 generation. However, this test method may be updated as new information and data are considered. For example, guidance on extending the F2 generation through reproduction may be potentially useful under certain circumstances (e.g., chemicals with high bioconcentration potential or indications of trans-generational effects in other taxa). This test method can be used to evaluate the potential chronic effects of chemicals, including potential endocrine disrupting chemicals, on fish. The method gives primary emphasis to potential population relevant effects (namely, adverse impacts on survival, development, growth and reproduction) for the calculation of a No Observed Effect Concentration (NOEC) or an Effect Concentration (ECx), although it should be noted that ECx approaches are rarely suitable for large studies of this type where increasing the number of test concentrations to allow for determination of the desired ECx may be impractical which may also cause significant animal welfare concerns due to the large number of animals used. For chemicals not requiring assessment over “multi-generations” or chemicals that are not potential endocrine disrupting chemicals, other test methods may be more appropriate (1). The Japanese medaka is the appropriate species for use in this test method, given its short life-cycle and the possibility to determine its genetic sex (2), which is considered a critical component in this test method. The specific methods and observational endpoints detailed in this method are applicable to Japanese medaka alone. Other small fish species (e.g., zebrafish) may be adapted to a similar test protocol.

2. This test method measures several biological endpoints. Primary emphasis is given to potential adverse effects on population relevant parameters including survival, gross development, growth and reproduction. Secondarily, in order to provide mechanistic information and provide linkage between results from other kinds of field and laboratory studies, where there is *a posteriori* evidence for a chemical having potential endocrine disrupter activity (*e.g.* androgenic or oestrogenic activity in other tests and assays) then other useful information is obtained by measuring *vitellogenin* (*vtg)* mRNA (or

vitellogenin protein, VTG), phenotypic secondary sex characteristics (SSC) as related to genetic sex, and evaluating histopathology. It should be noted that if a test chemical or its metabolites are not suspected of being EDCs, it may not be necessary to measure these secondary endpoints and less resource and animal intensive studies may be more appropriate (1). Definitions used in this test method are given in Appendix 1.

INITIAL CONSIDERATIONS AND LIMITATIONS

3. Due to the limited number of chemicals tested and laboratories involved in the validation of this rather complex assay, it is anticipated that when a sufficient number of studies is available to ascertain the impact of this new study design, the test method will be reviewed and if necessary revised in light of experience gained. The data can be used at Level 5 of the OECD Conceptual Framework for Testing and Assessment of Endocrine Disrupters (3). The test method begins by exposing adult fish (the F0 generation) to the test chemical during the reproduction phase. The exposure continues through development and reproduction in the F1 and hatch in the F2 generation; thus the assay allows evaluation of both structural and activational endocrine pathways. A weight of evidence approach may be undertaken when interpreting the endocrine related endpoints.

4. The test should include an adequate number of individuals to ensure sufficient power for the evaluation of reproduction-relevant endpoints (see Appendix 3) whilst ensuring that the number of animals used is the minimum required for animal welfare reasons. In view of the large numbers of test animals used, it is important to carefully consider the need for the test in relation to existing data which may already contain relevant information on many of the endpoints in the MEOGRT. Some assistance in this regard can be obtained from the OECD Fish Toxicity Testing Framework (1).

5. The test method has been designed primarily to distinguish the effects of a single substance. However, if a test on a mixture is required, then it should be considered whether it will provide acceptable results for the intended regulatory purpose.

6. Before beginning the test, it is important to have information about the physicochemical properties of the test chemical, particularly to allow the production of stable chemical solutions. It is also necessary to have an adequately sensitive analytical method for verifying test chemical concentrations.

PRINCIPLE OF THE TEST

7. The test is started by exposing sexually mature males and females (at least 12 wpf) in breeding pairs for 3 weeks, during which the test chemical is distributed in the organism of the parental generation (F0) according to its toxicokinetic behaviour. As near as possible to the first day of the fourth week, eggs are collected to start the F1 generation. During

rearing of the F1 generation (a total of 15 weeks), hatchability and survival are assessed. In addition, fish are sampled at 9-10 wpf for developmental endpoints and spawning is assessed for three weeks from 12 through 14 wpf. An F2 generation is started after the third week of the reproduction assessment and reared until completion of hatching.

TEST VALIDITY CRITERIA

8. The following criteria for test validity apply:

   - The dissolved oxygen concentration should be ≥ 60% of air saturation value throughout the test;
   - The mean water temperature over the entire duration of the study should be between 24 and 26°C. Brief excursions from the mean by individual aquaria should not be more than 2°C;
   - The mean fecundity of controls in each of the generations (F0 and F1) should be greater than 20 eggs per pair per day. Fertility of all the eggs produced during the assessment should be greater than 80%. In addition, 16 of the recommended 24 control breeding pairs (> 65%) should produce greater than 20 eggs per pair per day;
   - Hatchability of eggs should be ≥ 80% (average) in the controls (in each of the F1 and F2 generations);
   - Survival after hatching until 3 wpf and from 3 wpf through termination for the generation F1 (i.e. 15 wpf) should be ≥ 80% (average) and ≥ 90% (average), respectively in the controls (F1);
   - Evidence should be available to demonstrate that the concentrations of the test chemical in solution have been satisfactorily maintained within ± 20% of the mean measured values.

   Regarding water temperature, while not a validity criterion, replicates within a treatment should not be statistically different from each other, and treatment groups within the test should not be statistically different from each other (based on daily temperature measurements, and excluding brief excursions).

9. Although decreased reproduction may be observed in the higher exposure groups there should be sufficient reproduction in at least the third highest group and all lower groups of F0 to fill the hatching incubators. Furthermore, there should be adequate embryo survival in the third highest and lower exposure groups in F1 to allow endpoint evaluation at the sub-adult sampling (see paragraphs 36 and 38 and Appendix 9). Additionally, there should be at least minimal post-hatch survival (~20%) in the second highest exposure group of F1. These are not validity criteria, as such, but recommendations to permit robust NOECs to be calculated.

10. If a deviation from the test validity criteria is observed, the consequences should be considered in relation to the reliability of the test results and these deviations and considerations should be included in the test report.

DESCRIPTION OF THE METHOD

Apparatus

11. Normal laboratory equipment and especially the following:

(a) oxygen and pH meters;

(b) equipment for determination of water hardness and alkalinity;

(c) adequate apparatus for temperature control and preferably continuous monitoring;

(d) tanks made of chemically inert material and of a suitable capacity in relation to the recommended loading and stocking density (see Appendix 3);

(e) suitably accurate balance (i.e. accurate to ± 0.5 mg).

Water

12. Any water in which the test species shows suitable long-term survival and growth may be used as test water. It should be of constant quality during the period of the test. In order to ensure that the dilution water will not unduly influence the test result (for example by complexation of test chemical) or adversely affect the performance of the brood stock, samples should be taken at intervals for analysis. Measurements of heavy metals (e.g. Cu, Pb, Zn, Hg, Cd, Ni), major anions and cations (e.g. Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, Cl^{-}, SO_4^{2-}), pesticides, total organic carbon and suspended solids should be made, for example, every six months where a dilution water is known to be relatively constant in quality. Some chemical characteristics of acceptable dilution water are listed in Appendix 2. The pH of the water should be within the range 6.5 to 8.5, but during a given test it should be within a range of ± 0.5 pH units.

Exposure system

13. The design and materials used for the exposure system are not specified. Glass, stainless steel, or other chemically inert material should be used for construction of the test system that has not been contaminated during previous tests. For the purpose of this test, a well-suited exposure system may consist of a continuous flow-through system (4)(5)(6)(7)(8)(9)(10)(11)(12)(13).

Test solutions

14. Stock solution of the test chemical should be delivered into the exposure system by an appropriate pump. The flow rate of the stock solution should be calibrated in accordance with analytical confirmation of the test solutions before the initiation of exposure, and

checked volumetrically periodically during the test. The test solution in each chamber is renewed adequately (e.g., minimum of 5 volume renewals/day to up to 16 volume renewals/day or up to 20 ml/min flow) depending on the test chemical stability and water quality.

15. Test solutions of the chosen concentrations are prepared by dilution of a stock solution. The stock solution should preferably be prepared by simply mixing or agitating the test chemical in dilution water by mechanical means (e.g. stirring and/or ultra-sonication). Saturation columns/systems or passive dosing methods (14) can be used for achieving a suitably concentrated stock solution. All efforts should be made to avoid solvents or carriers because: (1) certain solvents themselves may result in toxicity and/or undesirable or unexpected responses, (2) testing chemicals above their water solubility (as can frequently occur through the use of solvents) can result in inaccurate determinations of effective concentrations, (3) the use of solvents in longer-term tests can result in a significant degree of "bio-filming" associated with microbial activity which may impact environmental conditions as well as the ability to maintain exposure concentrations and (4) in the absence of historical data that demonstrates that the solvent does not influence the outcome of the study, use of solvents requires a solvent control treatment which has animal welfare implications as additional animals are required to conduct the test. For difficult to test chemicals, a solvent may be employed as a last resort, and the OECD Guidance Document 23 on Aquatic Toxicity Testing of Difficult Substances and Mixtures (15) should be consulted to determine the best method. The choice of solvent will be determined by the chemical properties of the test chemical and the availability of historical data on use of the solvent. If solvent carriers are used, appropriate solvent controls should be evaluated in addition to non-solvent (negative) controls (dilution water only). In the event that use of a solvent is unavoidable, and microbial activity (bio-filming) occurs, recommend recording/reporting of the bio-filming per tank (at least weekly) throughout the test. Ideally, the solvent concentration should be kept constant in the solvent control and all test treatments. If the concentration of solvent is not kept constant, the highest concentration of solvent in the test treatment should be used in the solvent control. In cases where solvent carrier is used, maximum solvent concentrations should not exceed 100 µl/l or 100 mg/l (15), and it is recommended to keep solvent concentration as low as possible (*e.g.* < 20 µl/l) to avoid potential effect of the solvent on endpoints measured (16).

Test animals

Selection and holding of fish

16. The test species is Japanese medaka *Oryzias latipes* because of its short life-cycle and the possibility to determine genetic sex. Although other small fish species may be adapted to a similar test protocol, the specific methods and observational endpoints detailed in this test method are applicable to Japanese medaka alone (see paragraph 1). The medaka is readily induced to breed in captivity; published methods exist for its culture (17) (18) (19), and

data are available from short-term lethality, early life-stage and full life-cycle tests (5) (6) (8) (9) (20). All fish are maintained on a 16 h light:8 h dark photoperiod. The fish will be fed live brine shrimp, *Artemia* spp., nauplii which may be supplemented with a commercially available flake food if necessary. Commercially available flake food should be regularly analysed for contaminants.

17. As long as appropriate husbandry practices are followed, no specific culturing protocol is required. For example, medaka can be reared in 2 l tanks with 240 larval fish per tank until 4 wpf, then they can be reared in 2 l tanks with 10 fish per tank until 8 wpf, at which time, they transition to breeding pairs in 2 l tanks.

Acclimation and selection of fish

18. Test fish should be selected from a single laboratory stock which has been acclimated for at least two weeks prior to the test under conditions of water quality and illumination similar to those used in the test (Note: This acclimation period is not an *in situ* pre-exposure period). It is recommended that test fish be obtained from an in-house culture, as shipping of adult fish is stressful and may interfere with reliable spawning. Fish should be fed brine shrimp nauplii twice per day throughout the holding period and during the exposure phase, supplemented with a commercially available flake food if necessary. A minimum of 42 breeding pairs (54 breeding pairs if a solvent control is required due, in part, to lack of historical data to support the use of only the solvent control) are considered necessary to initiate this test to ensure adequate replication. In addition, each breeding pair of F0 should be verified to be XX-XY (i.e. normal complement of sex chromosomes in each sex) to avoid the possible inclusion of spontaneous XX males (see paragraph 39).

19. During the acclimation phase, mortalities in the culture fish should be recorded and the following criteria applied following a 48 h settling-down period:

- Mortalities of greater than 10% of the culture population in seven days preceding transfer to the test system: reject the entire batch;
- Mortalities of between 5% and 10% of the population in the seven days preceding transfer to the test system: acclimation for seven additional days to the 2-week acclimation period; if more than 5% mortality during the second seven days, reject the entire batch;
- Mortalities of less than 5% of the population in the seven days preceding transfer to the test system: accept the batch.

20. Fish should not receive treatment for disease in the two-week acclimation period preceding the test and during the exposure period, and disease treatment should be completely avoided if possible. Fish with clinical signs of disease should not be used in the study. A record of observations and any prophylactic and therapeutic disease treatments during the culture period preceding the test should be maintained.

21. The exposure phase should be started with sexually dimorphic, genetically sexed adult fish from a laboratory supply of reproductively mature animals cultured at 25 ± 2 °C. The fish should be identified as proven breeders (i.e. having produced viable offspring) during the week preceding exposure. For the whole group of fish used in the test, the range in individual weights by sex at the start of the test should be kept within ± 20% of the arithmetic mean weight of the same sex. A subsample of fish should be weighed before the test to estimate the mean weight. The fish selected should be at least 12 wpf, being a weight ≥ 300 mg for females and ≥ 250 mg for males.

TEST DESIGN

Test concentrations

22. It is recommended to use five chemical concentrations plus control(s). All sources of information should be considered when selecting the range of test concentrations, including quantitative structure activity relationships (QSARs), read-across from analogues, results of fish tests such as acute mortality assays (Chapter C.1 of this Annex), fish short-term reproduction assay (Chapter C.48 of this Annex) and other test methods e.g. Chapters C.15, C.37, C.41, C.47 or C.49 of this Annex (21) (22) (23) (24) (25) (26) if available, or if necessary, from a range-finding test possibly including a reproduction phase. If needed, the range-finding test may be conducted under conditions (water quality, test system, animal loading) similar to those used for the definitive test. If use of a solvent is necessary and no historical data are available, the range-finding test can be used to identify suitability of the solvent. The highest test concentration should not exceed the water solubility, 10 mg/l or $1/10^{th}$ of the 96h-LC50 (27). The lowest concentration should be a factor of 10- to 100-times lower than the highest concentration. The use of five concentrations in this test enables not only dose-response relationships to be measured, but also provides the Lowest Observed Effect Concentration (LOEC) and NOEC which are necessary for risk assessment in some regulatory programmes or jurisdictions. Generally, the spacing factor between nominal concentrations of the test chemical between adjacent treatment levels is ≤ 3.2.

Replicates within treatment groups and controls

23. A minimum of six replicate test chambers per test concentration should be used (see Appendix 7). During the reproductive phase (except F0 generation), replication structure is doubled for fecundity assessment and each replicate has only one breeding pair (see paragraph 42).

24. A dilution water control and, if needed, a solvent control should be run in addition to the test concentrations. A doubled number of replicate chambers for the controls should be used to ensure adequate statistical power (i.e., at least twelve replicates should be used for controls). During the reproductive phase, the number of replicates in the controls are

doubled (i.e. 24 replicates as a minimum and each replicate has only one mating pair). Following reproduction, control replicates should contain no more than 20 embryos (fish).

PROCEDURE

Initiation of test

25. The reproductively active adult fish used to start the F0 generation of the test are selected based on two criteria: age (typically more than 12 wpf but recommended not to exceed 16 wpf) and weight (should be ≥ 300 mg for females and ≥ 250 mg for males).

26. Female-male pairs that meet the above specifications are moved as individual pairs into each tank replicate, i.e. twelve replicates in controls and six replicates in chemical treatments at the initiation of the test. These tanks are randomly assigned a treatment (e.g., T1-T5 and control) and a replicate (e.g., A-L in controls and A-F in treatment), and then placed in the exposure system with the appropriate flow to each tank.

Conditions of exposure

27. A complete summary of test parameters and conditions can be found in Appendix 3. Adherence to these specifications should result in control fish with endpoint values similar to those listed in Appendix 4.

28. During the test, dissolved oxygen, pH, and temperature should be measured in at least one test vessel of each treatment group and the control. As a minimum, these measurements, except temperature, should be made once a week through the exposure period. The mean water temperature over the entire duration of the study should be between 24 and 26°C throughout the test. Temperature should be measured every day throughout the exposure period. The pH of the water should be within the range 6.5 to 8.5, but during a given test it should be within a range of ± 0.5 pH units. Replicates within a treatment should not be statistically different from each other, and treatment groups within the test should not be statistically different from each other (based on daily temperature measurements, and excluding brief excursions).

Duration of exposure

29. The test exposes sexually reproductive fish from F0 for three weeks. In week 4 on approximately test day 24, F1 is established and the F0 breeding pairs are humanely killed and weight and length are recorded (see Paragraph 34). This is followed by exposure of the F1 generation for 14 more weeks (total of 15 weeks for F1) and the F2 generation for two weeks until hatching. The total duration of the test is primarily 19 weeks (i.e., until F2 hatching). Timelines for the test are shown in Table 2 and further explained in detail in Appendix 9.

Feeding regime

30. Fish can be fed brine shrimp *Artemia* spp. (24-hours old nauplii) *ad libitum*, supplemented with a commercially available flake food if necessary. Commercially available flake food should be regularly analysed for contaminants such as organochlorine pesticides, polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs). Food with an elevated level of endocrine active substances (i.e., phytoestrogens) that could compromise the response of the test should be avoided. Uneaten food and faecal material should be removed from the test vessels as required, e.g. by carefully cleaning the bottom of each tank using a siphon. The sides and bottom of each tank should also be cleaned once or twice per week (e.g., by scraping with a spatula). An example of a feeding schedule can be found in Appendix 5. Feeding rate is based upon number of fish per replicate. Therefore, feeding rate is reduced if there are mortalities in a replicate.

Analytical determination and measurements

31. Prior to initiation of the exposure period, proper function of the chemical delivery system should be ensured. All analytical methods needed should be established, including sufficient knowledge of the chemical's stability in the test system. During the test, the concentrations of the test chemical are determined at appropriate intervals, preferably at least every week in one replicate for each treatment group, rotating between replicates of the same treatment group every week.

32. During the test, the flow rates of diluent and stock solution should be checked at intervals accordingly (e.g. at minimum three times a week). It is recommended that results be based on measured concentrations. However, if concentration of the test chemical in solution has been satisfactorily maintained within ± 20% of the mean measured values throughout the test, then the results can either be based on nominal or measured values. In case of chemicals that markedly accumulate in fish, the test concentrations may decrease as the fish grow. In such cases, it is recommended that the renewal rate of the test solution in each chamber be adapted to maintain test concentrations as constant as possible.

Observations and measured endpoints

33. Endpoints measured include fecundity, fertility, hatching, growth and survival for evaluation of possible population-level effects. Observations of behaviour should also be made daily, and any unusual behaviour noted. Other mechanistic endpoints include hepatic *vtg* mRNA or VTG protein levels by an immunoassay (28), sexual phenotypic markers such as characteristic male anal fin papillae, histological evaluation of gonadal sex, and histopathological evaluation of kidney, liver and gonad (see endpoint list in Table 1). All of these specific endpoints are evaluated in the context of a determination of the genetic sex of the individual, based on the presence or absence of the medaka male-sex determining gene *dmy* (see paragraph 41). Additionally, time to spawn is also evaluated. In addition, simple phenotypic sex ratios can be derived using the information from counts of anal fin papillae to define individual medaka as either phenotypically male or female. This

test method would not be expected to detect modest deviations from the expected sex ratio because the relatively small numbers of fish per replicate will not provide sufficient statistical power. Also, during the course of the histopathological assessment, the gonad is evaluated and much more powerful analyses for assessing the gonad phenotype in the context of the genetic sex are conducted.

34. The primary purpose of this test method is to assess the potential population relevant effects of a test chemical. Mechanistic endpoints (VTG, SSCs and certain gonadal histopathology effects) can also assist in determining whether any effect is mediated via endocrine activity. However, these mechanistic endpoints can also be influenced by systemic and other toxicities. Consequently, liver and kidney histopathology may also be assessed in detail to help better understand any responses in mechanistic endpoints. However, if these detailed evaluations are not performed, gross abnormalities observed incidentally during the histopathological evaluation should still be noted and reported.

Humane killing of fish

35. At termination of F0 and F1 generation exposure when sub-adult fish are subsampled, the fish should be euthanised with appropriate amounts of anaesthetic solution (e.g. Tricaine methane sulfonate, MS-222 (CAS.886-86-2), 100-500 mg/l) buffered with 300 mg/l $NaHCO_3$ (sodium bicarbonate, CAS.144-55-8) to reduce mucous membrane irritation. If fish are showing signs of considerable suffering (very severe and death can be reliably predicted) and considered moribund, animals should be anaesthetised and euthanised and treated as mortality for data analysis. When a fish is euthanised due to morbidity, this should be noted and reported. Depending on when the fish is euthanised during the study, retaining the fish for histopathology analysis may be conducted (fixing the fish for possible histopathology).

Handling of eggs and larval fish

Collection of eggs from breeding pairs to propagate the next generation

36. Egg collection is done on the first day (or first two days, if needed) of Test Week 4 to go from F0 to F1 and Test Week 18 to go from F1 to F2. Test Week 18 corresponds to F1, 15 wpf (weeks post fertilisation) adult fish. It is important that all eggs are removed from each tank the day before the egg collection starts to ensure all eggs collected from a breeding pair are from a single spawn. Following spawning, female medaka sometimes carry their eggs near the vent until the eggs can be deposited onto a substrate. With no substrate present in the tank, the eggs can be found either attached to the female or at the bottom of the tank. Depending on their location, eggs are either carefully removed from the female or siphoned from the bottom in Test Week 4 of F0 and Test Week 18 of F1. All eggs collected within a treatment are pooled prior to distribution to incubation chambers.

37. Egg filaments, which hold spawned eggs together, should be removed. Fertilised eggs (up to 20) are collected from each breeding pair (1 pair per replicate), are pooled by treatment, and systematically distributed to suitable incubation chambers (Appendix 6, 7). Using a good quality dissecting microscope, one can see hallmarks of early fertilisation/development such as raising of the fertilisation membrane (chorion), ongoing cell division, or formation of the blastula. The incubator chambers may be placed in separate "incubator aquaria" set up for each treatment (in which case water quality parameters and test chemical concentrations need to be measured in these), or in the replicate aquarium in which hatched larvae (*e.g.*, eleutheroembryo) will be contained. If a second day of collection (Test Day 23) is needed, all eggs from both days should be pooled and then systematically redistributed to each of the treatment replicates.

Rearing of eggs to hatching

38. Fertilised eggs are continually agitated *e.g.*, within the egg incubator by air bubbles or by vertically swinging the egg incubator. The mortalities of fertilised eggs (embryos) are checked and recorded daily. Dead eggs are removed from the incubators (Appendix 9). On the 7th day post fertilisation (dpf), the agitation is stopped or reduced so the fertilised eggs settle to the bottom of the incubator. This promotes hatching, typically over the next one or two days. For each treatment and control, hatchlings (young larvae; eleutheroembryo) are counted (pooled replicate basis). Fertilised eggs that have not hatched by twice the median day of hatch in the control (typically 16 or 18 dpf) are considered non-viable and discarded.

39. Twelve hatchlings are transferred into each replicate tank. The hatchlings from the incubation chambers are pooled and systematically distributed to replicate tanks (Appendix 7). This can be done by randomly selecting a hatchling from the treatment pool and sequentially adding a hatchling in an indiscriminate draw to a replicate aquarium. Each of the tanks should contain an equal number (n=12) of the hatched larvae (maximum 20 larvae each). If there are not enough hatchlings to fill all treatment replicates, then it is recommended to ensure as many replicates as possible have 12 hatchlings. Hatchlings can be handled safely with large-bore glass pipettes. Any additional hatchlings are humanely killed with anaesthetic. During the few weeks prior to the setup of breeding pairs, the day that the first spawning event is observed in each replicate should be recorded.

Setup of breeding pairs

Fin clipping and determination of genotypic sex

40. Determination of genotypic sex via fin clips is done at 9-10 wpf (i.e., Test Week 12-13 for F1 generation). All fish within a tank are anesthetised (using approved methods, e.g., IACUC) and a small tissue sample is taken from either the dorsal or the ventral tip of the caudal fin of each fish to determine the genotypic sex of the individual (29). The fish from a replicate can be housed in small cages, if possible one per cage, in the replicate tank.

Alternatively, two fish can be held in each cage if they are distinguishable from each other. One method is to differentially cut the caudal fin (*e.g.*, dorsal vs ventral tip) when taking the tissue sample.

41. The genotypic sex of medaka is determined by an identified and sequenced gene (*dmy*) which is located on the Y chromosome. The presence of *dmy* indicates a XY individual, regardless of phenotype, while the absence of *dmy* indicates a XX individual, regardless of phenotype (30); (31). Deoxyribose nucleic acid (DNA) from each fin clip is extracted and the presence or absence of *dmy* can be determined by polymerase chain reaction (PCR) methods (refer to Appendix 9 in Chapter C.41 of this Annex, or Appendix 3 and 4 in (29).

Establishment of breeding pairs

42. The information on genotypic sex is used to establish XX-XY breeding pairs regardless of external phenotype which may be altered by exposure to a test chemical. On the day after the genotypic sex of each fish is determined, two XX fish and two XY fish from each replicate are randomly selected and two XX-XY breeding pairs are established. If a replicate does not have either two XX or two XY fish, appropriate fish should be obtained from other replicates within the treatment. The priority is to have the recommended number of replicate breeding pairs (12) in each treatment and in the controls (24). Fish with obvious abnormalities (swim bladder problems, spinal deformities, extreme size variations, etc.) would be precluded when establishing breeding pairs. During the reproductive phase for F1 each replicate tank should contain only one breeding pair.

Sampling of sub-adults and endpoint assessment

Sampling of non-breeding pair fish

43. After the setup of breeding pairs, the fish not selected for further breeding are humanely killed for measurement of sub-adult endpoints in Test Week 12-13 (F1). It is extremely important that the fish are handled in such a way so that the genotypic sex determined for breeding pair selection can still be traced to an individual fish. All the data collected are analysed in the context of the genotypic sex of the specific fish. Each fish is used for a variety of endpoint measurements including: determination of survival rates of juvenile/sub-adult fish (Test Weeks 7-12/13 (F1), growth in length (standard length may be measured if the caudal fin has been shortened due to sampling for genetic sex analysis. Total length can be measured if only a portion of the caudal fin, dorsal or ventral, is sampled for *dmy*) and body mass (i.e., wet weight, blotted dry), liver *vtg* mRNA (or VTG) and anal fin papillae (see Tables 1 and 2). Note that weights and lengths of the breeding pairs are also required for calculating mean growth in a treatment group.

Tissue sampling and vitellogenin measurement

44. The liver is dissected, and should be stored at ≤ –70 °C until the *vtg* mRNA (or VTG) measurements. The tail of the fish, including the anal fin, is preserved in an appropriate

fixative (e.g. Davidson's) or photographed so that anal fin papillae can be counted at a later date. If desired, other tissues (i.e., gonad) may be sampled and preserved at this time). Liver VTG concentration should be quantified with a homologous ELISA technique (see the recommended procedures for medaka in Appendix 6 in Chapter C.48 of this Annex). Alternatively, the methods for vtg mRNA quantification, i.e., *vtg I* gene mRNA extraction from a liver sample and quantification of the number of copies of the *vtg I* gene (per ng of total mRNA) by quantitative PCR, have been established by the U.S EPA (29). Instead of determining the number of copies of the *vtg* gene in the control and treatment groups, a more resource friendly and less technically difficult method is to determine the relative (fold) change in *vtg I* expression from control and treatment groups.

Secondary sex characteristics

45. Under normal circumstances, only sexually mature male medaka have papillae, which develop on the joint plates of certain anal fin rays as a secondary sexual characteristic, providing a potential biomarker for endocrine disrupting effects. The method of counting anal fin papillae (the number of joint plates with papillae) is given in Appendix 8. Also the number of anal fin papillae per individual is used to categorise that individual as externally phenotypic male or female for the purpose of calculating a simple sex ratio per replicate. A medaka with any number greater than 0 is defined as a male; a medaka with 0 anal fin papillae is defined as a female.

Fecundity and fertility assessment

46. Fecundity and fertility are assessed in Test Weeks 1 through 3 in the F0 generation and Test Weeks 15 through 17 in the F1 generation. Eggs are collected daily from each breeding pair for 21 consecutive days. Eggs are gently removed from netted females and/or siphoned from the bottom of the aquarium each morning. Both fecundity and fertility are recorded daily for each replicate breeding pair. Fecundity is defined as the number of eggs spawned, and fertility is functionally defined as the number of fertilised and viable eggs at the time of counting. Counting should be done as soon as possible after egg collection.

47. Replicate fecundity is recorded daily as the number of eggs per breeding pair which is analysed by the recommended statistical procedures using the replicate means. Replicate fertility is the sum of the number of fertile eggs produced by a breeding pair divided by the sum of the number of eggs produced by that pair. Statistically fertility is analysed as a ratio per replicate. Replicate hatchability is the number of hatchlings divided by the number of embryos loaded (typically 20). Statistically hatchability is analysed as a ratio per replicate.

Sampling of adults and endpoint assessment

Sampling of breeding pair fish

48. Following Test Week 17 (i.e., after the F2 generation has successfully commenced), the F1 adults are humanely killed and various endpoints are assessed (see Tables 1 and 2). The

anal fin is imaged for assessing anal fin papillae (see Appendix 8), and/or the tail, just posterior to the vent, is removed and fixed for counting papillae later. A portion of the caudal fin may be sampled and archived at this time for verification of genetic sex (*dmy)* if desired. If needed, a tissue sample can be taken to repeat the *dmy* analysis to verify genetic sex of specific fish. The body cavity is opened to allow perfusion with appropriate fixatives (e.g., Davidson's) prior to submersing the entire body in the fixative. However, if an appropriate permeabilisation step is performed prior to fixation, the body cavity does not need to be opened.

Histopathology

49. Each fish is evaluated histologically for pathology in the gonadal tissue (30); (29). As referenced in paragraph 33, other mechanistic endpoints evaluated in this assay (i.e., VTG, SSCs and certain gonadal histopathology effects) may be influenced by systemic or other toxicities. Consequently, liver and kidney histopathology may also be assessed in detail to help better understand any responses in mechanistic endpoints. However, if these detailed evaluations are not performed, gross abnormalities observed incidentally during the histopathological evaluation should still be noted and reported. 'Reading down' from the highest treatment group (compared to the control) to a treatment with no effect could be considered, however, it is recommended to consult the histopathology guidance (29). Typically all samples are processed/sectioned after which are read by the pathologist. If using a 'read-down' approach, it is noted that the Rao-Scott Cochrane-Armitage by Slices (RSCABS) procedure uses the expectation that as dose levels increase the biological impact (the pathology) will increase as well. Therefore, one will lose power if only looking at a single high dose without any intermediate doses. If statistical analysis is not necessary to determine that the high dose has no effect, then this approach may be acceptable. The gonad phenotype is also derived from this evaluation

Other observations

50. The MEOGRT provides data that can be used (e.g., in a weight of evidence approach) to simultaneously evaluate at least two general types of adverse outcome pathways (AOPs) ending in reproductive impairment: (a) endocrine-mediated pathways involving disruption of the hypothalamus-pituitary-gonadal (HPG) endocrine axis; and, (b) pathways that cause reductions in survival, growth (length and weight), and reproduction through non-endocrine mediated toxicity. Endpoints typically measured in chronic toxicity tests such as the full life-cycle test and the early life-stage test are also included in this test and can be used to evaluate the hazards posed by both non-endocrine mediated toxic modes of action and endocrine-mediated toxicity pathways. During the test, observations of behaviour should be made daily, and any unusual behaviour should be noted. In addition, any mortality should be recorded and survival to the culling of fish (test week 6/7), survival after the culling to the sub-adult sampling (in 9-10 wpf), and survival from the pairing to the sampling of adult fish should be calculated.

Table 1: Endpoint overview of the MEOGRT*

Life-stage	Endpoint	Generation
Embryo (2 wpf)	Hatch (% and time to hatch)	F1, F2
Juvenile (4 wpf)	Survival	F1
Subadult (9 or10 wpf)	Survival	F1
	Growth (length and weight)	
	Vitellogenin (mRNA or protein)	
	Secondary sex characteristics (anal fin papillae)	
	External sex ratio	
	Time to 1st spawn	
Adult (12-14 wpf)	Reproduction (fecundity and fertility)	F0, F1
Adult (15 wpf)	Survival	F1
	Growth (length and weight)	
	Secondary sex characteristics (anal fin papillae)	
	Histopathology (gonad, liver, kidney)	

*These endpoints are to be statistically analysed

TIMELINE

51. A timeline for the MEOGRT illustrated in Table 2 shows the test. The MEOGRT includes 4 weeks of exposure to F0 adults and 15 weeks of exposure to the F1 generation, and exposure period for the second generation (F2), until hatching (2 wpf). Activity through the course of the MEOGRT is summarised in Appendix 9.

Table 2: Exposure and measurement endpoint timelines for the MEOGRT.

MEOGRT Exposure and Endpoint Timeline																			
F0	1	2	3	4															
F1				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
F2																		1	2
Test Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Lifestage Key					**Embryo**					**Larvae**					**Juvenile**			**Subadult**	**Adult**
Endpoints																			
Fecundity	F_0														F_1				
Fertility	F_0														F_1				
Hatch					F_1														F_2
Survival						F_1						F_1						F_1	
Growth				F_0								F_1						F_1	
Vitellogenin												F_1							
Secondary sex												F_1						F_1	
Histopathology																		F_1	
Test Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

- Experimental design has 7 groups of replicates
  - 5 for test chemical treatments
  - 2 for control treatments (4 if solvent is used)
- Within-group design
  - 12 replicates for reproduction, adult pathology and SSC (Wks 10 through to 18)
  - 6 replicates for hatch, survival, Vtg; and - subadult SSC and growth (Wks 1 through to 9)

SSC: secondary sex characters; Wks: weeks;
Vtg: vitellogenin

DATA REPORTING

Statistical analysis

52. Since genotypic sex is determined for all test fish, the data should be analysed for each genotypic sex separately (i.e., XY males and XX females). Failure to do this will greatly reduce the statistical power of any analysis. Statistical analyses of the data should preferably follow procedures described in the OECD document on Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application (32). Appendix 10 provides further guidance to the Statistical Analysis.

53. The test design and selection of statistical tests should permit adequate power to detect changes of biological importance in endpoints where a NOEC is to be reported (32). Reporting of relevant effect concentrations and parameters may depend upon the regulatory framework. The percent change in each endpoint that it is important to detect or estimate should be identified. The experimental design should be tailored to allow that. It is not likely that the same percent change applies to all endpoints, nor is it likely that a feasible experiment can be designed that will meet these criteria for all endpoints, so it is important to focus on the endpoints which are important for the respective experiment in designing the experiment appropriately. A statistical flow diagram and guidance is

available in Appendix 10 to help with the treatment of data and in the choice of the most appropriate statistical test or model to use. Other statistical approaches may be used, provided they are scientifically justified.

54. It will be necessary for variations to be analysed within each set of replicates using analysis of variance or contingency table procedures and sufficient appropriate statistical analysis methods used based on this analysis. In order to make a multiple comparison between the results at the individual concentrations and those for the controls, the step-down procedure (e.g., Jonckheere-Terpstra test) is recommended for continuous responses. Where the data are not consistent with a monotone concentration-response, Dunnett's test or Dunn's test should be used (after an adequate data transform, if necessary).

55. For fecundity, egg counts taken daily, but may be analysed as total egg counts or as a repeated measure. Appendix 10 provides the details of how this endpoint is analysed. For histopathology data which are in the form of severity scores, a new statistical test, Rao-Scott Cochran-Armitage by Slices (RSCABS), has been developed (33).

56. Any endpoints observed in chemical treatments that are significantly different from appropriate controls should be reported.

Data analysis considerations

Use of compromised treatment levels

57. Several factors are considered when determining whether a replicate or entire treatment demonstrates overt toxicity and should be removed from analysis. Overt toxicity is defined as >4 mortalities in any replicate between 3 wpf and 9 wpf that cannot be explained by technical error. Other signs of overt toxicity include haemorrhage, abnormal behaviours, abnormal swimming patterns, anorexia, and any other clinical signs of disease. For sub-lethal signs of toxicity, qualitative evaluations may be necessary, and should always be made in reference to the dilution water control group (clean water only). If overt toxicity is evident in the highest treatment(s), it is recommended that those treatments be censored from the analysis.

Solvent controls

58. The use of a solvent should only be considered as a last resort, when all other chemical delivery options have been considered. If a solvent is used, then a dilution water control should be run in concert. At the termination of the test, an evaluation of the potential effects of the solvent should be performed. This is done through a statistical comparison of the solvent control group and the dilution water control group. The most relevant endpoints for consideration in this analysis are growth determinants (weight), as these can be affected through generalised toxicities. If statistically significant differences are detected in these

endpoints between the dilution water control and solvent control groups, best professional judgment should be used to determine if the validity of the test is compromised. If the two controls differ, the treatments exposed to the chemical should be compared to the solvent control unless it is known that comparison to the dilution water control is preferred. If there is no statistically significant difference between the two control groups it is recommended that the treatments exposed to the test chemical are compared with the pooled (solvent and dilution-water control groups), unless it is known that comparison to either the dilution-water or solvent control group only is preferred.

Test report

59. The test report should include the following:

Test chemical: physical nature and, where relevant, physicochemical properties;

- Chemical identification data.

Mono-constituent substance:

- physical appearance, water solubility, and additional relevant physicochemical properties;
- chemical identification, such as IUPAC or CAS name, CAS number, SMILES or InChI code, structural formula, purity, chemical identity of impurities as appropriate and practically feasible, etc. (including the organic carbon content, if appropriate).

Multi-constituent substance, UVCBs and mixtures:

- characterised as far as possible by chemical identity (see above), quantitative occurrence and relevant physicochemical properties of the constituents.

Test species:

- Scientific name, strain if available, source and method of harvesting of the fertilised eggs and subsequent handling.

Test conditions:

- Photoperiod(s);
- Test design (e.g. chamber size, material and water volume, number of test chambers and replicates, number of hatchlings per replicates);
- Method of preparation of stock solutions and frequency of renewal (the solubilising agent and its concentration should be given, when used);
- Method of dosing the test chemical (*e.g.* pumps, diluting systems);
- The recovery efficiency of the method and the nominal test concentrations, the limit of quantification, the means of the measured values and their standard deviations in the test

vessels and the method by which these were attained and evidence that the measurements refer to the concentrations of the test chemical in true solution;

- Dilution water characteristics: pH, hardness, temperature, dissolved oxygen concentration, residual chlorine levels (if measured), total organic carbon (if measured), suspended solids (if measured), salinity of the test medium (if measured) and any other measurements made;
- The nominal test concentrations, the means of the measured values and their standard deviations;
- Water quality within test vessels, pH, temperature (daily) and dissolved oxygen concentration;
- Detailed information on feeding (e.g. type of foods, source, amount given and frequency).

Results:

- Evidence that controls met the overall validation criteria;
- Data for the control (plus solvent control when used) and the treatment groups as follows, hatching (hatchability and time to hatch) for F1 and F2, post hatch survival for F1, growth (length and body weight) for F1, genotypic sex and sexual differentiation (e.g. secondary sex characteristics based on anal fin papillae and gonadal histology) for F1, phenotypic sex for F1, secondary sex characteristics (anal fin papillae) for F1 *vtg* mRNA (or VTG protein) for F1, histopathology assessment (gonad, liver and kidney) for F1 and reproduction (fecundity and fertility) for F0, F1; (see Tables 1 and 2).
- Approach for the statistical analysis (regression analysis or analysis of the variance) and treatment of data (statistical tests and models used);
- No observed effect concentration (NOEC) for each response assessed;
- Lowest observed effect concentration (LOEC) for each response assessed (at $p = 0.05$); EC_x for each response assessed, if applicable, and confidence intervals (e.g. 90% or 95%) and a graph of the fitted model used for its calculation, the slope of the concentration-response curve, the formula of the regression model, the estimated model parameters and their standard errors.
- Any deviation from this test method and deviations from the acceptance criteria, and considerations of potential consequences on the outcome of the test.

60. For the results of endpoint measurements, mean values and their standard deviations (on both replicate and concentration basis, if possible) should be presented.

LITERATURE

(1) OECD (2012a). Fish Toxicity Testing Framework, Environment, Health and Safety Publications, Series on Testing and Assessment (No. 171), Organisation for Economic Cooperation and Development, Paris.

(1) Padilla S, Cowden J, Hinton DE, Yuen B, Law S, Kullman SW, Johnson R, Hardman RC, Flynn K and Au DWT. (2009). Use of Medaka in Toxicity Testing. Current Protocols in Toxicology 39: 1-36.

(2) OECD (2012b). Guidance Document on Standardised Test Guidelines for Evaluating Endocrine Disrupters. Environment, Health and Safety Publications, Series on Testing and Assessment (No. 150), Organisation for Economic Cooperation and Development, Paris.

(3) Benoit DA, Mattson VR, Olson DL. (1982). A Continuous-Flow Mini-Diluter System for Toxicity Testing. Water Research 16: 457–464.

(4) Yokota H, Tsuruda Y, Maeda M, Oshima Y, Tadokoro H, Nakazono A, Honjo T and Kobayashi K. (2000). Effect of Bisphenol A on the Early Life Stage in Japanese Medaka (*Oryzias Latipes*). Environmental Toxicology and Chemistry 19: 1925-1930.

(5) Yokota H, Seki M, Maeda M, Oshima Y, Tadokoro H, Honjo T and Kobayashi K. (2001). Life-Cycle Toxicity of 4-Nonylphenol to Medaka (*Oryzias Latipes*). Environmental Toxicology and Chemistry 20: 2552-2560.

(6) Kang IJ, Yokota H, Oshima Y, Tsuruda Y, Yamaguchi T, Maeda M, Imada N, Tadokoro H and Honjo T. (2002). Effects of 17β-Estradiol on the Reproduction of Japanese Medaka (*Oryzias Latipes*). Chemosphere 47: 71–80.

(7) Seki M, Yokota H, Matsubara H, Tsuruda Y, Maeda M, Tadokoro H and Kobayashi K. (2002). Effect of Ethinylestradiol on the Reproduction and Induction of Vitellogenin and Testis-Ova in Medaka (*Oryzias Latipes*). Environmental Toxicology and Chemistry 21: 1692-1698.

(8) Seki M, Yokota H, Matsubara H, Maeda M, Tadokoro H and Kobayashi K. (2003). Fish Full Life-Cycle Testing for the Weak Estrogen 4-T*ert*-Pentylphenol on Medaka (*Oryzias Latipes*). Environmental Toxicology and Chemistry 22: 1487-1496.

(9) Hirai N, Nanba A, Koshio M, Kondo T, Morita M and Tatarazako N. (2006a). Feminization of Japanese Medaka (*Oryzias latipes*) Exposed to 17β-Estradiol: Effect of Exposure Period on Spawning Performance in Sex-Transformed Females. Aquatic Toxicology 79: 288-295.

(10) Hirai N, Nanba A, Koshio M, Kondo T, Morita M and Tatarazako N. (2006b). Feminization of Japanese Medaka (*Oryzias latipes*) Exposed to 17β-Estradiol: Formation of Testis-Ova and Sex-Transformation During Early-Ontogeny. Aquatic Toxicology 77: 78-86.

(11) Nakamaura A, Tamura I, Takanobu H, Yamamuro M, Iguchi T and Tatarazako N. (2015). Fish Multigeneration Test with Preliminary Short-Term Reproduction Assay for Estrone Using Japanese Medaka (*Oryzias Latipes).* Journal of Applied Toxicology 35:11-23.

(12) U.S. Environmental Protection Agency (2013). Validation of the Medaka Multigeneration Test: Integrated Summary Report. Available at: http://www.epa.gov/scipoly/sap/meetings/2013/062513meeting.html.

(13) Adolfsson-Erici M, Åkerman G, Jahnke A, Mayer P and McLachlan M. (2012). A Flow-Through Passive Dosing System for Continuously Supplying Aqueous Solutions of Hydrophobic Chemicals to Bioconcentration and Aquatic Toxicity Tests. Chemosphere 86: 593-599.

(14) OECD (2000). Guidance Document on Aquatic Toxicity Testing of Difficult Substances and Mixtures. OECD Environment, Health and Safety Publications, Series on Testing and Assessment (No. 23.), Organisation for Economic Cooperation and Development, Paris.

(15) Hutchinson TH., Shillabeer N., Winter MJ and Pickford DB. (2006). Acute and Chronic Effects of Carrier Solvents in Aquatic Organisms: A Critical Review. Review. Aquatic Toxicology 76: 69–92.

(16) Denny JS, Spehar RL, Mead KE and Yousuff SC. (1991). Guidelines for Culturing the Japanese Medaka, *Oryzias latipes*. US EPA/600/3-91/064.

(17) Koger CS, Teh SJ and Hinton DE. (1999). Variations of Light and Temperature Regimes and Resulting Effects on Reproductive Parameters in Medaka (*Oryzias Latipes*). Biology of Reproduction 61: 1287-1293.

(18) Kinoshita M, Murata K, Naruse K and Tanaka M. (2009). Medaka: Biology, Management, and Experimental Protocols, Wiley- Blackwell.

(19) Gormley K and Teather K. (2003). Developmental, Behavioral, and Reproductive Effects Experienced by Japanese Medaka in Response to Short-Term Exposure to Endosulfan. Ecotoxicology and Environmental Safety 54: 330-338.

(20) Chapter C.15 of this Annex, Fish, Short-term Toxicity Test on Embryo and Sac-fry Stages.

(21) Chapter C.37 of this Annex, 21-day Fish Assay: A Short-Term Screening for Oestrogenic and Androgenic Activity, and Aromatase Inhibition.

(22) Chapter C.41 of this Annex, Fish Sexual Development Test.

(23) Chapter C.48 of this Annex, Fish Short Term Reproduction Assay.

(24) Chapter C.47 of this Annex, Fish, Early-life Stage Toxicity Test.

(25) Chapter C.49 of this Annex, Fish Embryo Acute Toxicity (FET) Test.

(26) Wheeler JR, Panter GH, Weltje L and Thorpe KL. (2013). Test Concentration Setting for Fish *In Viv*o Endocrine Screening Assays. Chemosphere 92: 1067-1076.

(27) Tatarazako N, Koshio M, Hori H, Morita M and Iguchi T. (2004). Validation of an Enzyme-Linked Immunosorbent Assay Method for Vitellogenin in the Medaka. Journal of Health Science 50: 301-308.

(28) OECD (2015). Guidance Document on Medaka Histopathology Techniques and Evaluation. Environment, Health and Safety Publications, Series on Testing and Assessment (No. 227). Organisation for Economic Cooperation and Development, Paris.

(29) Nanda I, Hornung U, Kondo M, Schmid M and Schartl M. (2003). Common Spontaneous Sex-Reversed XX Males of the Medaka *Oryzias Latipes*. Genetics 163: 245–251.

(30) Shinomiya, A, Otake H. Togashi K. Hamaguchi S and Sakaizumi M. (2004). Field Survey of Sex-Reversals in the Medaka, *Oryzias Latipes*: Genotypic Sexing of Wild Populations, Zoological Science 21: 613-619.

(31) OECD (2014). Current Approaches in the Statistical Analysis of Ecotoxicity Data: A guidance to application (annexes to this publication exist as a separate document), OECD Publishing, Paris.

(32) Green JW, Springer TA, Saulnier AN and Swintek J. (2014). Statistical Analysis of Histopathology Endpoints. Environmental Toxicology and Chemistry 33: 1108-1116.

Appendix 1

DEFINITIONS

Chemical: A substance or a mixture.

ELISA: Enzyme-Linked Immunosorbent Assay

Fecundity = number of eggs;

Fertility = number of viable eggs/fecundity;

Fork length (FL) refers to the length from the tip of the snout to the end of the middle caudal fin rays and is used in fishes in which it is difficult to tell where the vertebral column ends www.fishbase.org

Hatchability = hatchlings/number of embryos loaded into an incubator

IACUC: Institutional Animal Care and Use Committee

Standard length (SL) refers to the length of a fish measured from the tip of the snout to the posterior end of the last vertebra or to the posterior end of the midlateral portion of the hypural plate. Simply put, this measurement excludes the length of the caudal fin. (www.fishbase.org)

Total length (TL) refers to the length from the tip of the snout to the tip of the longer lobe of the caudal fin, usually measured with the lobes compressed along the midline. It is a straight-line measure, not measured over the curve of the body (www.fishbase.org)

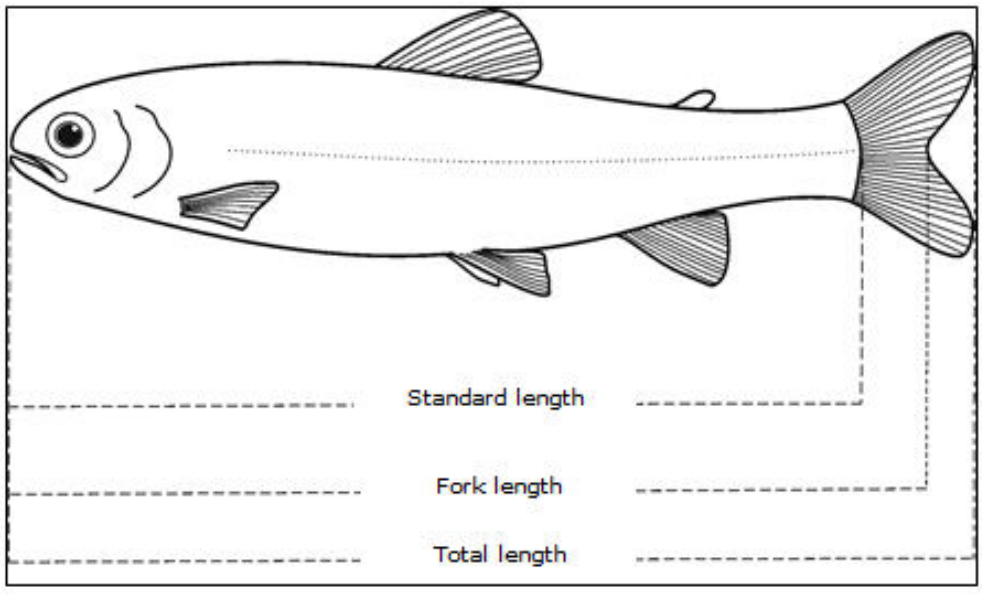


Figure 1: Description of the different lengths, used

ECx: (Effect concentration for x% effect) is the concentration that causes an x% of an effect on test organisms within a given exposure period when compared with a control. For example, an EC50 is a concentration estimated to cause an effect on a test end point in 50% of an exposed population over a defined exposure period.

Flow-through test is a test with continued flow of test solutions through the test system during the duration of exposure.

HPG axis: hypothalamic-pituitary-gonadal axis.

IUPAC: International Union of Pure and Applied Chemistry.

Loading rate: The wet weight of fish per volume of water.

Lowest observed effect concentration (LOEC) is the lowest tested concentration of a test chemical at which the chemical is observed to have a statistically significant effect (at $p < 0.05$) when compared with the control. However, all test concentrations above the LOEC should have a harmful effect equal to or greater than those observed at the LOEC. When these two conditions cannot be satisfied, a full explanation should be given for how the LOEC (and hence the NOEC) has been selected. Appendix 5 and 6 provide guidance.

Median Lethal Concentration (LC50): is the concentration of a test chemical that is estimated to be lethal to 50% of the test organisms within the test duration.

No observed effect concentration (NOEC) is the test concentration immediately below the LOEC, which when compared with the control, has no statistically significant effect ($p < 0.05$), within a stated exposure period.

SMILES: Simplified Molecular Input Line Entry Specification.

Stocking density: The number of fish per volume of water.

Test chemical: Any substance or mixture tested using this test method.

UVCB: Substances of unknown or variable composition, complex reaction products or biological materials.

VTG: Vitellogenin is a phospholipoglycoprotein precursor to egg yolk protein that normally occurs in sexually active females of all oviparous species.

WPF: Weeks post fertilisation

Appendix 2

SOME CHEMICAL CHARACTERISTICS OF AN ACCEPTABLE DILUTION WATER

Substance	Limit concentration
Particulate matter	5 mg/l
Total organic carbon	2 mg/l
Un-ionised ammonia	1 µg/l
Residual chlorine	10 µg/l
Total organophosphorous pesticides	50 ng/l
Total organochlorine pesticides plus polychlorinated biphenyls	50 ng/l
Total organic chlorine	25 ng/l
Aluminium	1 µg/l
Arsenic	1 µg/l
Chromium	1 µg/l
Cobalt	1 µg/l
Copper	1 µg/l
Iron	1 µg/l
Lead	1 µg/l
Nickel	1 µg/l
Zinc	1 µg/l
Cadmium	100 ng/l
Mercury	100 ng/l
Silver	100 ng/l

Appendix 3

TEST CONDITIONS FOR THE MEOGRT

1. Recommended species	Japanese medaka (*Oryzias latipes*)
2. Test type	Continuous flow-through
3. Water temperature	The nominal test temperature is 25°C. The mean temperature throughout the test in each tank is 24-26 °C.
4. Illumination quality	Fluorescent bulbs (wide spectrum and ~150 lumens/m^2) (~150 lux).
	16 h light:8 h dark
6. Loading rate	F0: 2 adults/replicate; F1: initiated with maximum 20 eggs (embryos)/replicate, reduced to 12 embryos/replicate at hatch then 2 adults (XX-XY breeding pair) at 9-10 wpf for reproductive phase
7. Minimum test chamber usable volume	1.8 l (e.g., test chamber size: 18x9x15 cm)
8. Volume exchanges of test solutions	Minimum of 5 volume renewal/day to up to 16 volume renewal/day (or 20 ml/min flow)
9. Age of test organisms at initiation	F0: > 12 wpf but recommended not to exceed 16 wpf
10. Number of organisms per replicate	F0: 2 fish (male and female pair); F1: maximum 20 fish (eggs)/replicate (produced from F0 and F1 breeding pairs).
11. Number of treatments	5 test chemical treatments plus appropriate control(s)
12. Number of replicates per treatment	Minimum 6 replicates per treatment for test chemical and minimum 12 replicates for control, and for solvent control, if used (the number of replicates are doubled within reproduction phase in F1)
13. Number of organisms per test	Minimum of 84 fish in F0 and 504 in F1. (If solvent control is used, then 108 fish in F0 and 648 fish in F1). The unit counted is the post-eleutheroembryo.
14. Feeding regime	Fish are fed brine shrimp, *Artemia* spp., (24-hour old nauplii) *ad libitum*, supplemented with a commercially available flake food if needed (An example feeding schedule to ensure adequate growth and development to support robust reproduction can be found in Appendix 6).
15. Aeration	None unless dissolved oxygen approaches <60 % of air saturation value
16. Dilution water	Clean surface, well or reconstituted water or dechlorinated tap water.

17. Exposure period	Primarily 19 weeks (from F0 to F2 hatching)
18. Biological endpoints (primary)	Hatchability (F1 and F2); survival (F1, from hatch to 4 wpf (end of larval/beginning of juvenile), from 4 to 9 (or 10) wpf (beginning of juvenile to subadult) and from 9 to 15 wpf (subadult to adult termination)); growth (F1, length and weight at 9 and 15 wpf); secondary sex characteristics (F1, anal fin papillae at 9 and 15 wpf); vitellogenin (F1, *vtg* mRNA or VTG protein at 15wpf); phenotypic sex (F1, via gonad histology at 15 wpf); reproduction (F0 and F1, fecundity and fertility for 21 days); time to spawn (F1); and histopathology (F1, gonad, liver and kidney at 15 wpf)
19. Test validity criteria	Dissolved oxygen of $\geq$ 60% air saturation value; mean water temperature of 24-26^{o}C throughout the test; successful reproduction of $\geq$ 65% females in control(s); mean daily fecundity of $\geq$ 20 eggs in control(s); hatchability of $\geq$ 80 % (average) in the controls (in each of the F1 and F2); survival after hatching until 3 wpf of $\geq$ 80 % (average) and from 3 wpf through termination for the generation of $\geq$ 90 % (average) in the controls (F1), concentrations of the test chemical in solution should be satisfactorily maintained within ± 20% of the mean measured values.

Appendix 4

GUIDANCE ON TYPICAL CONTROL VALUES

It should be noted that these control values are based on a limited number of validation studies, and may be subject to amendment in the light of further experience.

Growth

Weight and length measurements are taken for all fish sampled at 9 (or 10) and 15 weeks post fertilisation (wpf). Following this protocol will yield expected wet weights at 9 wpf of 85-145 mg for males and 95-150 mg for females. The expected weights at 15 wpf are 250-330 mg for males and 280-350 mg for females. While there may be substantial deviations from these ranges for individual fish, control mean weights substantially outside of these ranges, especially lower, would suggest problems with feeding, temperature control, water quality, disease or any combination of these factors.

Hatch

Hatching success in controls is typically around 90%, however, values as low as 80% are not uncommon. Hatch success less than 75% may indicate insufficient agitation of the developing eggs or inadequate care in handling the eggs such as lack of timely removal of dead eggs leading to fungal infestation.

Survival

Survival rates until 3 wpf from hatch and after 3 wpf are usually 90% or greater for controls but survival rates in early life stages as low as 80% for controls are not alarming. Survival rates in controls of less than 80% would be cause for concern and may indicate insufficient cleaning of the aquaria leading to loss of larval fish through disease or from suffocation due to low dissolved oxygen levels. Mortality may also occur as a result of injury during tank cleaning and by the loss of larval fish to the drain system of the tank.

Vitellogenin gene

While absolute levels of *vitellogenin* (*vtg*) gene, expressed as copies/ng of total mRNA, may vary greatly between laboratories due to the procedures or instrumentation used, the ratio of *vtg* should be around 200 times greater in control females versus control males. It is not uncommon for this ratio to be as high as from 1000 to 2000, however, ratios less than 200 are suspect and may indicate problems with sample contamination or problems with the procedure and/or reagents used.

Secondary sex characteristics

For males, the normal range of Secondary Sex Characteristics, defined as the total number of segments in the fin-rays of the anal fin papillae, is 40-80 segments at 9-10 wpf. By 15 wpf, the range for control males should be about 80-120 and 0 for control females. For unexplained reasons, in rare instances some males have no papillae present by 9 wpf but

since all control males develop papillae by 15 wpf, this is most likely caused by delayed development. The presence of papillae in control females indicates the presence of XX males in the population.

XX-males

The normal background incidence of XX males in culture appears to be about 4 % or less at 25 ºC with the incidence increasing with increased temperature. Steps should be taken to minimise the proportion of XX males in the population. Since the incidence of XX males appears to have a genetic component and is therefore heritable, monitoring the culture stock and ensuring that XX males are not used to propagate the culture stock is an effective means to reduce the incidence of XX males in the population.

Spawning activity

Spawning activity in the control replicates should be monitored daily prior to conducting the fecundity assessment. The control pairs can be qualitatively assessed visually for evidence of spawning activity. By12-14 wpf most control pairs should be spawning. Low numbers of spawning pairs by this time indicates potential problems with the health, maturity or well-being of the fish.

Fecundity

Healthy, well fed 12-14 wpf medaka generally spawn daily, producing in the range of 15 to 50 eggs per day. Egg production for 16 of the recommended 24 control breeding pairs (> 65%) should produce greater than 20 eggs per pair per day and may reach as high as about 40 eggs per day. Less than this amount may indicate immature, malnourished or unhealthy spawning pairs.

Fertility

The percentage of fertile eggs for control spawning pairs is typically in the 90% range with values in the mid-to-upper 90s not uncommon. Fertility rates of less than 80% for control eggs are suspect and may indicate either unhealthy individuals or less than ideal culture conditions.

Appendix 5

AN EXAMPLE OF A FEEDING SCHEDULE

An example of a feeding schedule to ensure adequate growth and development to support robust reproduction is shown in Table 1. Deviations from this feeding schedule may be acceptable, but it is recommended that they are tested to verify that acceptable growth and reproduction be observed. In order to follow the suggested feeding schedule, the dry weight of brine shrimp per volume of brine shrimp slurry needs to be determined prior to starting the test. This can be done by weighing a defined volume of brine shrimp slurry that has been dried for 24 hours at 60 °C on pre-weighed pans. To account for the weight of the salts in the slurry, an identical volume of the same salt solution used in the slurry should also be dried, weighed, and subtracted from the dried brine shrimp slurry weight. Alternatively, the brine shrimp can be filtered and rinsed with distilled water before drying, thereby eliminating the need to measure the weight of a "salt blank". This information is used to convert the information in the Table from dry weight of brine shrimp to volume of brine shrimp slurry to be fed per fish. In addition, it is recommended that aliquots of the brine shrimp slurry are weighed weekly to verify the correct dry weight of brine shrimp being fed.

Table 1: Example of a feeding schedule

Time (post-hatch)	Brine Shrimp (mg dry weight/fish/day)
Day 1	0.5
Day 2	0.5
Day 3	0.6
Day 4	0.7
Day 5	0.8
Day 6	1.0
Day 7	1.3
Day 8	1.7
Day 9	2.2
Day 10	2.8
Day 11	3.5
Day 12	4.2
Day 13	4.5
Day 14	4.8
Day 15	5.2
Day 16-21	5.6
Week 4	7.7
Week 5	9.0
Week 6	11.0
Week 7	13.5
Week 8-sacrifice	22.5

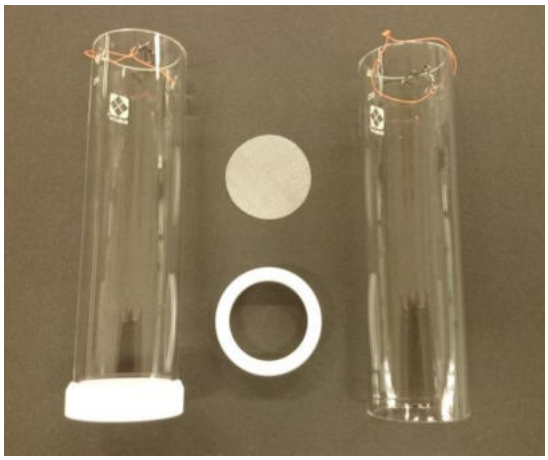
Appendix 6

EXAMPLES OF AN EGG INCUBATION CHAMBER

Example A

This incubator consists of a transected glass centrifuge tube, connected by a stainless steel sleeve and held in place by the centrifuge screw top cap. A small glass or stainless steel tube projects through the cap and is positioned near the rounded bottom, gently bubbling air to suspend the eggs and reducing between-egg transmission of saprophytic fungal infections while also facilitating chemical exchange between the incubator and the holding tank.

Example B

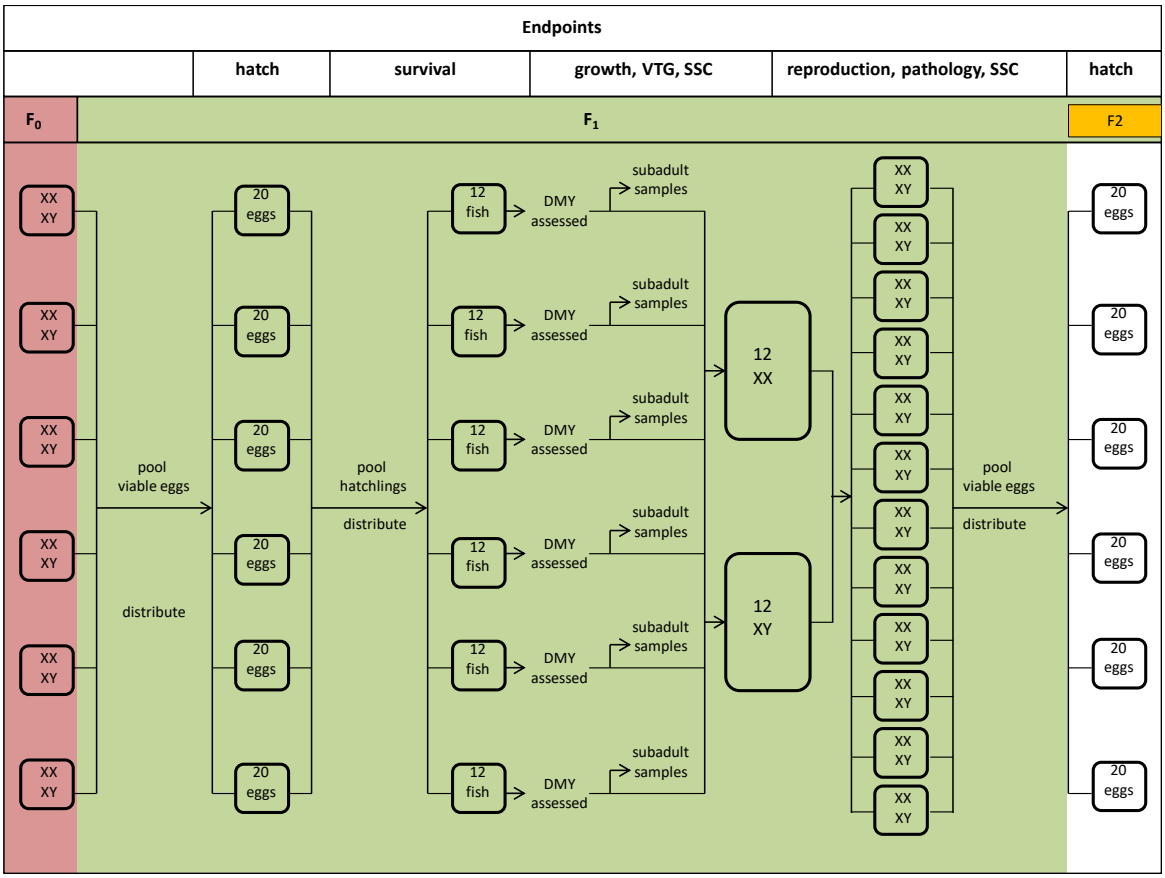


This incubator consists of a glass cylinder body (5 cm diameter and 10 cm height) and stainless wire mesh (0.25 φ and 32 mesh) which is attached to the bottom of the body with a PTFE ring. The incubators are suspended from the lifting bar to tanks, and shaken vertically (approximately 5 cm amplitude) in an appropriate cycle (approximately once every 4 seconds) for medaka eggs.

Appendix 7

SCHEMATIC DIAGRAM FOR POOLING AND POPULATING REPLICATES THROUGHOUT THE MEOGRT TEST METHOD

Figure 1: Pooling and repopulating replicates throughout the MEOGRT. The figure represents one treatment or ½ of a control. Due to pooling, replicate identity is not continuous throughout the test. Note that the term 'eggs' refers to viable, fertilised eggs (equivalent to embryos).



Treatments and Replication.

The test method recommends five test chemical treatments using technical grade material and a negative control. The number of replicates per treatment does not remain constant throughout the MEOGRT, and the number of replicates in the control treatment is double of any single test chemical treatment. In F0, each test chemical treatment has six replicates while the negative control treatment has 12 replicates. Solvents are highly discouraged, and if used, a justification for both the use of a solvent and the choice of solvent should be included in the MEOGRT report. Also, if a solvent is used, two types of controls are necessary: a) a solvent control, and b) a negative control. These two control groups should each consist of a full complement of replicates at all points within the MEOGRT timeline. Throughout test organism development in the F1 generation (and F2, until hatch), this replicate structure remains the same. However, in the adult stage when F1 breeding pairs are setup, the number of reproducing pair replicates per treatment is optimally doubled;

therefore there are up to 12 replicate pairs in each test chemical treatment and 24 replicate pairs in the control group (and another 24 replicate pairs in the solvent control, if needed). The determination of hatch from embryos spawned by the F1 pairs is done on the same replicate structure as was done for the embryos spawned by the F0 pairs, meaning initially six replicates per test chemical treatment and 12 replicates in the control group(s).

Appendix 8

COUNTING ANAL FIN PAPILLAE

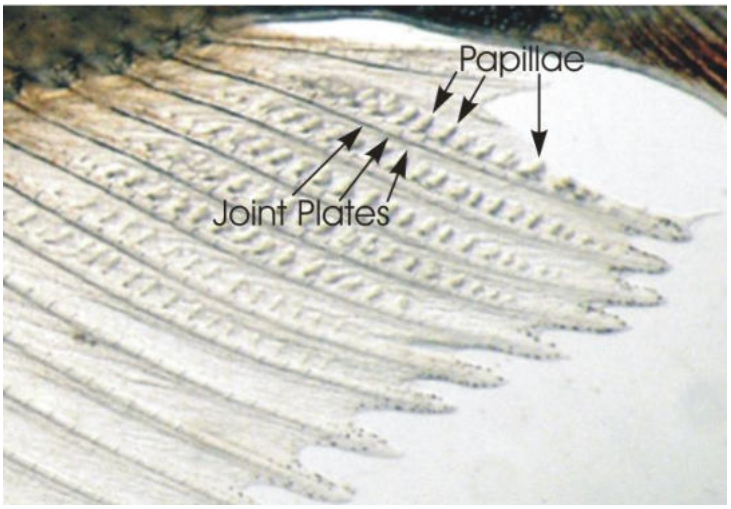
Major Materials and Reagents

- Dissecting microscope (with optional camera attached)
- Fixative (e.g., Davidson's (Bouin's is not recommended)), if not counting from image

Procedures

After necropsy, the anal fin should be imaged to allow for convenient counting of anal fin papillae. While imaging is the recommended method, the anal fin can be fixed with Davidson's fixative or other appropriate fixative for approximately 1 minute. It is important to keep the anal fin flat during fixation to allow for easier counting of papillae. The carcass with the anal fin can be stored in Davidson's fixative or other appropriate fixative until analysed. Count the number of joint plates (see **Figure 1**) with papillae which protrude from the posterior margin of the joint plate.

Figure 1: Anal fin papillae



Appendix 9

DETAILED TIMELINE OF MEOGRT

Test Weeks 1-3 (F0)

The F0 generation spawning fish that have met the selection criteria (see para. 16-20) are exposed for three weeks to allow the developing gametes and gonadal tissues to be exposed to the test chemical. Each replicate tank houses a single breeding fish pair (XX female-XY male breeding pair). Spawned eggs are collected, counted and assessed for fertility for 21 consecutive days, starting at Test Day 1.

Test Week 4 (F0 and F1)

It is preferable that the fertilised and viable eggs (embryos) are collected on a single day; however, if there are not enough embryos, the embryos may be collected over two days If collected over two days, all embryos within the treatments that were collected on the first day are pooled with those collected on the second day. Then the total pooled embryos for each treatment are randomly distributed to each of the replicate incubators at 20 embryos per incubator. The mortalities of fertilised eggs (embryos) are checked and recorded daily. Dead eggs are removed from the incubators (death in fertilised eggs may be denoted by, particularly in the early stages, a marked loss of translucency and change in colouration, caused by coagulation and/or precipitation of protein, leading to a white opaque appearance; OECD 2010).

Note: If a single treatment requires a second day of collection, all treatments (including controls) need to follow this procedure. If after the second day of collection there are inadequate numbers of embryos within a treatment to load 20 embryos per incubator, then reduce the number of embryos loaded within that specific treatment to 15 embryos per incubator. If there are not enough embryos to load 15 per incubator, then reduce the number of replicate incubators until there are enough embryos for 15 per incubator. Additionally, more breeding pairs per treatment and controls could be added in F0 to produce more eggs to reach the recommended 20 per replicate.

On Test Day 24, the F0 breeding pairs are humanely killed and weight and length are recorded. If necessary F0 breeding pairs maybe kept for an additional 1-2 days in order to restart F1.

Test Weeks 5-6 (F1)

One to two days before the anticipated start of hatching, stop or reduce the agitation of the incubating eggs to expedite hatching. As embryos hatch on each day, hatchlings are pooled by treatment and systematically distributed to each replicate larval tank within a specific treatment with no more than 12 hatchlings. This is done by randomly selecting hatchlings and placing a single hatchling in successive replicates in an indiscriminate draw, moving in order through the specific treatment replicates until all replicates within the treatment have 12 hatchlings. If there are not enough hatchlings to fill all replicates then ensure as many

replicates as possible have 12 hatchlings to start the F1 phase.

The eggs that have not hatched by twice the median control day of hatch are considered non-viable and discarded. The number of hatchlings is recorded and hatching success (hatchability) is calculated in each replicate.

Test Weeks 7-11 (F1)

The survival of larval fish is checked and recorded daily in all replicates. On Test Day 43, the number of surviving fish in each replicate is recorded as well as the initial number of hatchlings placed in the replicate (nominally twelve). This allows for the calculation of the percent survival from hatch to the subadult stage.

Test Weeks (F1)

On Test Day 78-85, a small sample is taken from the caudal fin of each fish to determine the genotypic sex of the individual (i.e., fin clipping) for all fish. This information is used to establish breeding pairs.

Within three days after the genotypic sex of each fish is determined, 12 breeding pairs per treatment and 24 pairs per control are randomly established. Two XX and XY fish from each replicate are randomly selected and then pooled by sex, and then randomly selected to establish a breeding pair (i.e., XX-XY pair). A minimum 12 replicates per chemical treatment and minimum 24 replicates for the control are established with one breeding pair per replicate. If a replicate does not have either two XX or two XY fish available for pooling, then fish with the appropriate gender genotype should be obtained from other replicates within the treatment.

The remaining fish (maximum 8 fish per replicate) are humanely killed and sampled for the various subadult endpoints. The *dmy* data (XX or XY) for all the subadult samples are retained to ensure that all endpoint data can be related to the genetic sex of each individual fish.

Test Weeks 13-14 (F1)

The exposure continues as the subadult breeding pairs develop into adults. On Test Day 98 (i.e. the day before egg collection is started), eggs are removed from both the aquaria and the females.

Test Weeks 15-17 (F1)

Spawned eggs are collected daily for 21 consecutive days in each replicate, and assessed for fecundity and fertility.

Test Week 18 (repeat of Test Week 4) (F1 and F2)

On Test Day 120, eggs collection is done in each replicate tank in the morning. The collected eggs are assessed and fertilised eggs (filaments removed) from each of the

breeding pairs are pooled by treatment, and systematically distributed to egg incubation chambers with 20 fertilised eggs per incubator. The incubators may be placed in separate "incubator tanks" set up for each treatment or in the replicate tank that upon hatch will contain the hatched larvae. It is preferable that the embryos are collected on a single day; however, if there are not enough embryos, the embryos may be collected over two days. If collected over two days, all embryos within the treatments that were collected on the first day are pooled with those collected on the second day. Then the total pooled embryos for each treatment are randomly distributed to each of the replicate incubators at 20 embryos per incubator. Note: If a single treatment requires a second day of collection, all treatments (including controls) need to follow this procedure. If after the second day of collection there is inadequate numbers of embryos within a treatment to load 20 embryos per incubator, reduce the number of embryos loaded within that specific treatment to 15 embryos per incubator. If there are not enough embryos to load 15 per incubator, reduce the number of replicate incubators until there are enough embryos for 15 per incubator.

On Test Day 121 (or Test Day 122, to ensure the F2 has started well), the F1 breeding pairs are humanely killed and analysed for the adult endpoints. If necessary F1 breeding pairs maybe kept for an additional 1-2 days in order to restart F2.

Test Weeks 19-20 (F2)

One to two days before the anticipated start of hatching, stop or reduce the agitation of the incubating eggs to expedite hatching. If the test is terminated by the completion of the F2 hatching, each day the hatchlings are tallied and discarded. (Embryos that have not hatched after a prolonged incubation time, defined as twice the median control day of hatch, are considered non-viable).

Appendix 10

STATISTICAL ANALYSIS

The types of biological data generated in the MEOGRT are not unique to it and except for pathology data, many appropriate statistical methodologies have been developed to properly analyse similar data depending on the characteristics of the data including normality, variance homogeneity, whether the study design lends itself to hypothesis testing or regression analysis, parametric versus non-parametric tests, etc. In general principle, the suggested statistical analyses follow the recommendations of the OECD for ecotoxicity data (OECD 2006) and a decision flowchart for MEOGRT data analysis can be seen in Figure 2.

It is assumed that most often the datasets will display monotonic responses. Additionally, the issue of using a one-tailed statistical test versus a two-tailed statistical test should be considered. Unless there is a biological reasoning that would make a one-tailed test inappropriate, it is suggested that one-tailed tests be used. While the following section recommends certain statistical tests, if more appropriate and/or powerful statistical methods are developed for application to the specific data generated in the MEOGRT, those statistical tests would be used to leverage those advantages.

The MEOGRT data should be analysed separately for each genotypic sex. There are two strategies to analysing the data from sex reversed fish (either XX males or XY females). 1) Censor all data from sex reversed fish across the entire test except the prevalence of sex reversal in each replicate. 2) Leave the data from all sex reversed fish in the dataset and analyse based upon genotype.

Histopathology data

Histopathology data are reported as severity scores which are evaluated using a newly developed statistical procedure, the Rao-Scott Cochrane-Armitage by Slices (RSCABS), (Green *et al*., 2014). The *Rao-Scott* adjustment retains test-replication information; the *by Slices* procedure incorporates the biological expectation that severity scores tend to increase with increasing treatment concentrations. For each diagnosis, the RSCABS output specifies which treatments have higher prevalence of pathology than controls and the associated severity level.

Fecundity data

Analyses for fecundity data consist of a step-down Jonckheere-Terpstra or Williams' test to determine treatment effects, provided the data are consistent with a monotone concentration-response. With a step-down test, all comparisons are done at the 0.05 significance level and no adjustment for the number of comparisons made. The data are expected to be consistent with a monotone concentration response, but this can be verified either by visual inspection of the data or by constructing linear and quadratic contrasts of treatment means after a rank-order transform of the data. Unless the quadratic contrast is significant and the linear contrast is not significant, the trend test is done. Otherwise, Dunnett's test is used to determine treatment effects if the data are normally distributed with

homogeneous variances. If those requirements are not met, then Dunn's test with a Bonferonni-Holm adjustment is used. All indicated tests are done independently of any overall F- or Kruskal-Wallis test. Further details are provided in OECD 2006.

Alternative methods can be used, such as a generalised linear model with Poisson errors for egg counts (with no transform), if justified statistically (Cameron and Trividi, 2013). Statistical advice is recommended if an alternative approach is used.

Daily Egg Count within a Single Generation

The ANOVA model is given by Y=Time*Time+Treatment + *Treatment + Time*Treatment + *Time*Treatment, with random effects of Replicate(Generation*Treatment), and Time*Replicate(Treatment), allowing for unequal variance components of both types across generations. Here Time refers to the frequency of egg counts (e.g., Day or Week). This is a repeated measures analysis, with the correlations between observations on the same replicates accounting for the repeated measures nature of the data.

Main effects of treatment are tested using the Dunnett (or Dunnett-Hsu) test, which adjusts for the number of comparisons. Adjustments for the main effect of generation or time are needed, for with these two factors, there is no "control" level and every pair of levels is a comparison of possible interest. For these two main effects, if the F-test for the main effect is significant at the 0.05 level, then the pairwise comparisons across levels of that factor can then be tested at the 0.05 level without further adjustment.

The model includes two- and three-factor interactions, so that a main effect for, say, time, may not be significant even though time has a significant impact on results. Thus, if a two- or three-factor interaction involving time is significant at the 0.05 level, then one can accept the comparisons of levels of time at the 0.05 significance level without further adjustment.

Next are F-tests for significance of treatment within time, the so-called slices in the ANOVA table. If, for example, the slice for treatment within F1 and time 12, is significant at the 0.05 level, then the pairwise comparisons for treatment within F1 and time 12 can be accepted at the 0.05 level without further adjustment. Similar statements apply to tests for time within F1 and treatment and for generation within a time and treatment.

Finally, for comparisons not falling under any of the above categories, comparisons should be adjusted using the Bonferroni-Holm adjustment to p-values. Further information on analyses of such models can be found in Hocking (1985) and Hochberg and Tamhane (1987).

Alternatively, the raw data are recorded and presented in the study report as the fecundity (number of eggs) per replicate for each day. The replicate mean of the raw data should be calculated then a square root transformation applied. A one-way ANOVA on the transformed replicate means should be calculated followed by Dunnett contrasts. It may also be helpful to visually inspect the fecundity data of each treatment and/or replicate with a scatterplot that displays the data through time. This will allow an informal assessment of potential effects through time.

All other biological data

The statistical analyses are based on the underlying assumption that with proper dose selection the data will be monotonic. Thus, data are assumed to be monotonic and they are formally evaluated for monotonicity by using linear and quadratic contrasts. If the data are monotonic, a Jonckheere-Terpstra on replicate medians trend test (as advised in OECD 2006) is recommended. If the quadratic contrast is significant and the linear contrast is not, the data are considered non-monotonic.

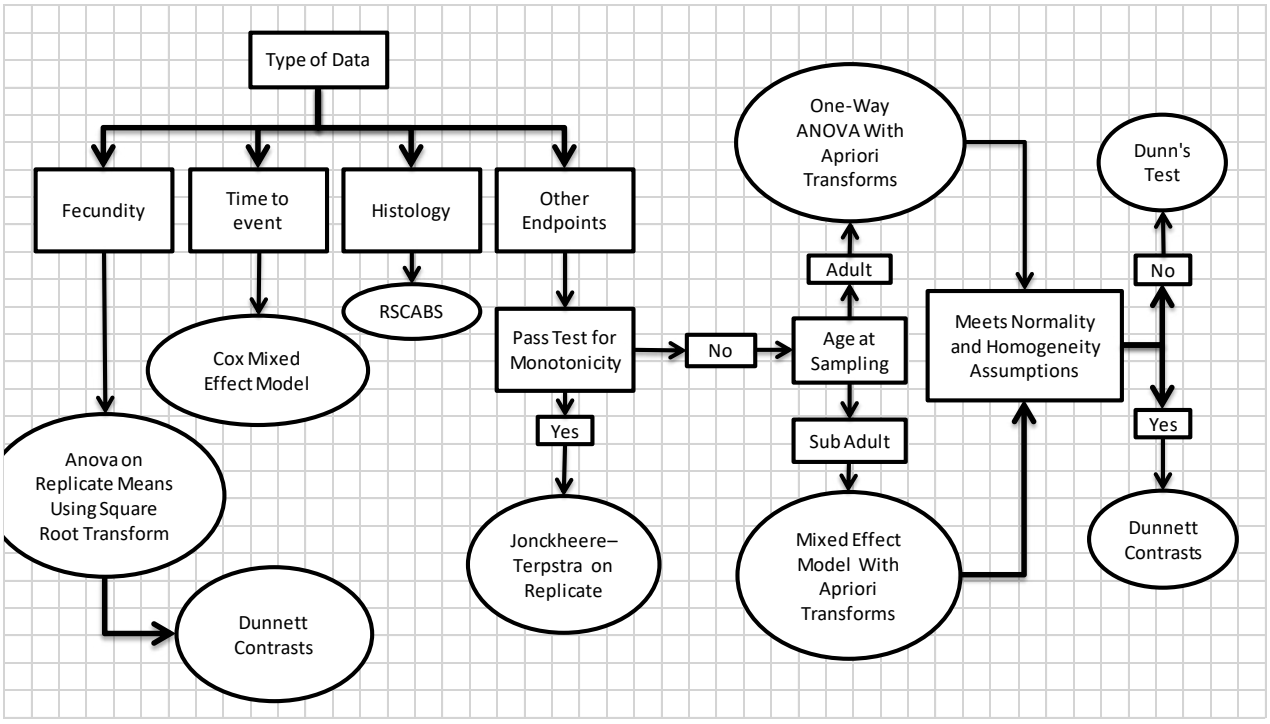
If the data are non-monotonic, in particular because of the reduced response of the highest one or two treatments, consideration should be given to censoring the dataset so that the analysis is done without those treatments. This decision will need to be made with professional judgment and all available data, especially data that indicates overt toxicity at those treatment levels.

For weight and length, no transforms are recommended although they may occasionally be necessary. However, a log transformation is recommended for the vitellogenin data; a square root transformation is recommended for the SSC data (anal fin papillae); an arcsine-square root transformation is recommended for the data on proportion hatching, percent survival, sex ratio, and percent fertile eggs. Time to hatch and time to first spawn should be treated as time to event data, with individual embryos not hatching in the defined period or replicates never spawning treated as right-censored data. Time to hatch should be calculated from the median day of hatch of each replicate. These endpoints should be analysed using a mixed-effects Cox proportional hazard model.

The biological data from adult samples has one measurement per replicate, that is, there are one XX fish and one XY fish per replicate aquarium. Therefore, it is recommended that a one-way ANOVA be done on the replicate means. If the assumptions of the ANOVA (normality and variance homogeneity as assessed on the residuals of the ANOVA by Shapiro-Wilks test and Levene's test, respectively) are met, Dunnett contrasts should be used to determine treatments that were different from the control. On the other hand, if the assumptions of the ANOVA are not met, then a Dunn's test should be done to determine which treatments were different from control. A similar procedure is recommended for data that are in the form of percentages (fertility, hatch, and survival).

The biological data from subadult samples has from 1 to 8 measurements per replicate, that is, there can be variable numbers of individuals that contribute to the replicate mean for each genotypic sex. Therefore, it is recommended that a mixed effects ANOVA model be used followed by Dunnett contrasts, if the normality and variance homogeneity assumptions were met (on the residuals of the mixed effects ANOVA). If they were not met, then a Dunn's test should be done to determine which treatments were different from control.

Figure 2: Flow chart for the recommended statistical procedures for MEOGRT data analysis.



Literature

(1) OECD (2014). Current Approaches in the Statistical Analysis of Ecotoxicity Data: A guidance to application (annexes to this publication exist as a separate document), OECD Publishing, Paris.

(2) Cameron AC and Trivedi PK (2013). Regression Analysis of Count Data, 2nd edition, Econometric Society Monograph No 53, Cambridge University Press.

(3) Hocking RR (1985). The Analysis of Linear Models, Monterey, CA: Brooks/Cole.

(4) Hochberg Y and Tamhane AC (1987). Multiple Comparison Procedures. John Wiley and Sons, New York.

C.53 THE LARVAL AMPHIBIAN GROWTH AND DEVELOPMENT ASSAY (LAGDA)

INTRODUCTION

1. This test method is equivalent to OECD test guideline 241 (2015). The need to develop and validate an assay capable of identifying and characterising the adverse consequences of exposure to toxic chemicals in amphibians, originates from concerns that environmental levels of chemicals may cause adverse effects in both humans and wildlife. The OECD test guideline of the Larval Amphibian Growth and Development Assay (LAGDA) describes a toxicity test with an amphibian species that considers growth and development from fertilisation through the early juvenile period. It is an assay (typically 16 weeks) that assesses early development, metamorphosis, survival, growth, and partial reproductive maturation. It also enables measurement of a suite of other endpoints that allows for diagnostic evaluation of suspected endocrine disrupting chemicals (EDCs) or other types of developmental and reproductive toxicants. The method described in this test method is derived from validation work on African clawed frog (*Xenopus laevis*) by the U.S. Environmental Protection Agency (U.S. EPA) with supporting work in Japan (1). Although other amphibian species may be adapted to a growth and developmental test protocol with ability to determine genetic sex being an important component, the specific methods and observational endpoints detailed in this test method are applicable to *Xenopus laevis* alone.

2. The LAGDA serves as a higher tier test with an amphibian for collecting more comprehensive concentration-response information on adverse effects suitable for use in hazard identification and characterisation, and in ecological risk assessment. The assay fits at Level 4 of the OECD Conceptual Framework for Testing and Assessment of Endocrine Disrupters, where *in vivo* assays also provide data on adverse effects on endocrine relevant endpoints (2). The general experimental design entails exposing *X. laevis* embryos at Nieuwkoop and Faber (NF) stage 8-10 (3) to a minimum of four different concentrations of test chemical (generally spaced at not less than half-logarithmic intervals) and control(s) until 10 weeks after the median time to NF stage 62 in the control, with one interim sub-sample at NF stage 62 ($\leq$ 45 post fertilisation; usually around 45 days (dpf). There are four replicates in each test concentration with eight replicates for the control. Endpoints evaluated during the course of the exposure (at the interim sub-sample and final sample at completion of the test) include those indicative of generalised toxicity: mortality, abnormal behaviour, and growth determinations (length and weight), as well as endpoints designed to characterise specific endocrine toxicity modes of action targeting oestrogen, androgen or thyroid-mediated physiological processes. The method gives primary emphasis to potential population relevant effects (namely, adverse impacts on survival, development, growth and reproductive development) for the calculation of a No Observed Effect Concentration (NOEC) or an Effect Concentration causing x% change (ECx) in the endpoint measured.

Although it should be noted that ECx approaches are rarely suitable for large studies of this type where increasing the number of test concentrations to allow for determination of the desired ECx may be impractical. It should also be noted that the method does not cover the reproductive phase itself. Definitions used in this test method are given in Appendix 1.

INITIAL CONSIDERATIONS AND LIMITATIONS

3. Due to the limited number of chemicals tested and laboratories involved in the validation of this rather complex assay, especially inter-laboratory reproducibility is not documented with experimental data so far, it is anticipated that when a sufficient number of studies is available to ascertain the impact of this new study design, OECD test guideline 241 will be reviewed and if necessary revised in light of experience gained. The LAGDA is an important assay to address potential contributors to amphibian population declines by evaluating the effects from exposure to chemicals during the sensitive larval stage, where effects on survival and development, including normal development of reproductive organs, may adversely affect populations.

4. The test is designed to detect an apical effect(s) resulting from both endocrine and non-endocrine mechanisms, and includes diagnostic endpoints which are partly specific to key endocrine modalities. It should be noted that until the LAGDA was developed, no validated assay existed that served this function for amphibians.

5. Before beginning the assay, it is important to have information about the physicochemical properties of the test chemical, particularly to allow the production of stable chemical solutions. It is also necessary to have an adequately sensitive analytical method for verifying test chemical concentrations. Over a duration of approximate 16 weeks, the assay requires a total number of 480 animals, i.e., *X. laevis* embryos, (or 640 embryos, if a solvent control is used) to ensure sufficient power of the test for the evaluation of population-relevant endpoints such as growth, development and reproductive maturation.

6. Before use of the test method for regulatory testing of a mixture, it should be considered whether it will provide acceptable results for the intended regulatory purpose. Furthermore, this assay does not evaluate fecundity directly, so it may not be applicable for use at a more advanced stage than Level 4 of the OECD Conceptual Framework for Testing and Assessment of Endocrine Disrupters.

SCIENTIFIC BASIS FOR THE TEST METHOD

7. Much of our current understanding of amphibian biology has been obtained using the laboratory model species *X. laevis.* This species can be routinely cultured in the laboratory, ovulation can be induced using human chorionic gonadotropin (hCG) and animal stocks are readily available from commercial breeders.

8. Like all vertebrates, reproduction in amphibians is under the control of the hypothalamic pituitary gonadal (HPG) axis (4). Oestrogens and androgens are mediators of this endocrine system, directing the development and physiology of sexually-dimorphic tissues. There are three distinct phases in the life cycle of amphibians when this axis is especially active: (1) gonadal differentiation during larval development, (2) development of secondary sex characteristics and gonadal maturation during the juvenile phase and (3) functional reproduction of adults. Each of these three developmental windows are likely susceptible to endocrine perturbation by certain chemicals such as estrogens and androgens, ultimately leading to a loss of reproductive fitness by the organisms.

9. The gonads begin development at NF stage 43, when the bipotential genital ridge first develops. Differentiation of the gonads begins at NF stage 52 when primordial germ cells either migrate to medullary tissue (males) or remain in the cortical region (females) of the developing gonads (3). This process of sexual differentiation of the gonads was first reported to be susceptible to chemical alteration in *Xenopus* in the 1950's (5) (6). Exposure of tadpoles to estradiol during this period of gonad differentiation results in sex reversal of males that when raised to adulthood are fully functional females (7) (8). Functional sex reversal of females into males is also possible and has been reported following implantation of testis tissue in tadpoles (9). However, although exposure to an aromatase inhibitor also causes functional sex reversal in *X. tropicalis* (10), this has not been shown to occur in *X. laevis*. Historically, toxicant effects on gonadal differentiation have been assessed by histological examination of the gonads at metamorphosis and sex reversal could only be determined by analysis of sex ratios. Until recently, there had been no means to directly determine the genetic sex of *Xenopus*. However, recent establishment of sex linked markers in *X. laevis* make it possible to determine genetic sex and allows for the direct identification of sex reversed animals (11).

10. In males, juvenile development proceeds as blood levels of testosterone increase corresponding with the development of secondary sex characteristics as well as testis development. In females, estradiol is produced by the ovaries resulting in the appearance of vitellogenin (VTG) in the plasma, vitellogenic oocytes in the ovary and the development of oviducts (12). Oviducts are female secondary sex characteristics that function in oocyte maturation during reproduction. Jelly coats are applied to the outside of oocytes as they pass through the oviduct and collect in the ovisac, ready for fertilisation. Oviduct development appears to be regulated by oestrogens as development correlates with blood estradiol levels in *X. laevis* (13) and *X. tropicalis* (12). The development of oviducts in males following exposure to polychlorinated biphenyl compounds (14) and 4-*tert*-octylphenol (15) has been reported.

PRINCIPLE OF THE TEST

11. The test design entails exposing *X. laevis* embryos at NF stage 8-10 via the water route to four different concentrations of test chemical as well as control(s) until 10 weeks after the median time to NF stage 62 in the control with one interim sub-sample at NF stage 62. While it may also be possible to dose highly hydrophobic chemicals via the feed, there has been little experience using this exposure route in this assay to date. There are four replicates in each test concentration with eight replicates for each control used. Endpoints evaluated during the course of the exposure include those indicative of generalised toxicity (i.e., mortality, abnormal behaviour and growth determinations (length and weight)), as well as endpoints designed to characterise specific endocrine toxicity modes of action targeting oestrogen-, androgen-, or thyroid-mediated physiological processes (i.e. thyroid histopathology, gonad and gonad duct histopathology, abnormal development, plasma vitellogenin (optional), and genotypic/phenotypic sex ratios) .

TEST VALIDITY CRITERIA

12. The following criteria for test validity apply:

- The dissolved oxygen concentration should be ≥ 40% of air saturation value throughout the test;
- The water temperature should be in the range of 21 ± 1 °C and the inter-replicate and the inter-treatment differentials should not exceed 1.0 ºC;
- pH of the test solution should be maintained between 6.5 and 8.5, and the inter-replicate and the inter-treatment differentials should not exceed 0.5;
- Evidence should be available to demonstrate that the concentrations of the test chemical in solution have been satisfactorily maintained within ± 20% of the mean measured values;
- Mortality over the exposure period should be ≤ 20% in each replicate in the controls;
- ≥ 70% viability in the spawn chosen to start the study;
- The median time to NF stage 62 of the controls should be ≤ 45 days.
- The mean weight of test organisms at NF stage 62 and at the termination of the assay in controls and solvent controls (if used) should reach 1.0 ± 0.2 and 11.5 ± 3 g, respectively.

13. While not a validity criterion, it is recommended that at least three treatment levels with three uncompromised replicates be available for analysis. Excessive mortality, which compromises a treatment, is defined as > 4 mortalities (> 20%) in 2 or more replicates that cannot be explained by technical error. At least three treatment levels without obvious overt toxicity should be available for analysis. Signs of overt toxicity may include, but are not limited to, floating on the surface, lying on the bottom of the tank, inverted or irregular swimming, lack of surfacing activity, and being nonresponsive to stimuli, morphological abnormalities (*e.g.*, limb deformities), hemorrhagic lesions, and abdominal oedema.

14. In case a deviation from the test validity criteria is observed, the consequences should be considered in relation to the reliability of the test results, and these deviations and considerations should be included in the test report.

DESCRIPTION OF THE METHODS

Apparatus

15. Normal laboratory equipment and especially the following:

    (a) temperature controlling apparatus (e.g., heaters or coolers adjustable to 21 ± 1 ºC);

    (b) thermometer;

    (c) binocular dissection microscope and dissection tools;

    (d) digital camera with at least 4 megapixel resolution and micro function (if needed);

    (e) analytical balance capable of measuring to 0.001 mg or 1 µg;

    (f) dissolved oxygen meter and pH meter;

    (g) light intensity meter capable of measuring in lux units.

Water

Source and quality

16. Any dilution water that is locally available (e.g. spring water or charcoal-filtered tap water) and permits normal growth and development of *X. laevis* can be used, and evidence of normal growth in this water should be available. Because local water quality can differ substantially from one area to another, analysis of water quality should be undertaken, particularly if historical data on the utility of the water for raising amphibian larvae is not available. Measurements of heavy metals (e.g. Cu, Pb, Zn, Hg, Cd, Ni), major anions and cations (e.g. Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, Cl^{-}, SO_4^{2-}), pesticides, total organic carbon and suspended solids should be made before testing begins and/or, for example, every six months where a dilution water is known to be relatively constant in quality. Some chemical characteristics of acceptable dilution water are listed in Appendix 2.

Iodide concentration in test water

17. In order for the thyroid gland to synthesise thyroid hormones to support normal metamorphosis, sufficient iodide should to be available to the larvae through a combination of aqueous and dietary sources. Currently, there are no empirically derived guidelines for minimum iodide concentrations in either food or water to ensure proper development. However, iodide availability may affect the responsiveness of the thyroid system to thyroid active agents and is known to modulate the basal activity of the thyroid gland which deserves attention when interpreting the results from thyroid histopathology. Based on previous work, successful performance of the assay has been demonstrated when

dilution water iodide (I^-) concentrations range between 0.5 and 10 µg/l. Ideally, the minimum iodide concentration in the dilution water throughout the test should be 0.5 µg/l (added as the sodium or potassium salt). If the test water is reconstituted from deionised water, iodine should be added at a minimum concentration of 0.5 µg/l. The measured iodide concentrations from the test water (i.e., dilution water) and the supplementation of the test water with iodine or other salts (if used) should be reported. Iodine content may also be measured in food(s) in addition to test water.

Exposure system

18. The test was developed using a flow-through diluter system. The system components should have water-contact components of glass, stainless steel, and/or other chemically inert materials. Exposure tanks should be glass or stainless steel aquaria and tank usable volume should be between 4.0 and 10.0 l (minimum water depth of 10 to 15 cm). The system should be capable of supporting all exposure concentrations, a control, and a solvent control, if necessary, with four replicates per treatment and eight in the controls. The flow rate to each tank should be constant in consideration of both the maintenance of biological conditions and chemical exposure. It is recommended that flow rates should be appropriate (*e.g.*, at least 5 tank turnovers per day) to avoid chemical concentration declines due to metabolism by both the test organisms and aquatic microorganisms present in the aquaria or abiotic routes of degradation (hydrolysis, photolysis) or dissipation (volatilisation, sorption). The treatment tanks should be randomly assigned to a position in the exposure system to reduce potential positional effects, including slight variations in temperature, light intensity, *etc*. Further information on setting up flow-through exposure systems can be obtained from the ASTM Standard Guide for Conducting Acute Toxicity Tests on Test Materials with Fishes, Macroinvertebrates, and Amphibians (16).

Chemical delivery: preparation of test solutions

19. To make test solutions in the exposure system, stock solution of the test chemical should be dosed into the exposure system by an appropriate pump or other apparatus. The flow rate of the stock solution should be calibrated in accordance with analytical confirmation of the test solutions before the initiation of exposure, and checked volumetrically periodically during the test. The test solution in each chamber should be renewed at a minimum of 5 volume renewals/day.

20. The method used to introduce the test chemical to the system can vary depending on its physicochemical properties. Therefore, prior to the test, baseline information about the chemical that is relevant to determining its testability should be obtained. Useful information about test chemical-specific properties include the structural formula, molecular weight, purity, stability in water and light, pK_a and K_{ow}, water solubility (preferably in the test medium) and vapour pressure as well as results of a test for ready biodegradability (test method C.4 (17) or C.29 (18)). Solubility and vapour pressure can be

used to calculate Henry's law constant, which will indicate whether losses due to evaporation of the test chemical may occur. Conduct of this test without the information listed above should be carefully considered as the study design will be dependent on the physicochemical properties of the test chemical and, without these data test results may be difficult to interpret or meaningless. A reliable analytical method for the quantification of the test chemical in the test solutions with known and reported accuracy and limit of detection should be available. Water soluble test chemicals can be dissolved in aliquots of dilution water at a concentration which allows delivery at the target test concentration in a flow-through system. Chemicals which are liquid or solid at room temperature and moderately soluble in water may require liquid:liquid or liquid:solid (*e.g.*, glass wool column) saturators (19). While it may also be possible to dose very hydrophobic test chemicals via the feed, there has been little experience using that exposure route in this assay.

21. Test solutions of the chosen concentrations are prepared by dilution of a stock solution. The stock solution should preferably be prepared by simply mixing or agitating the test chemical in dilution water by mechanical means (e.g. stirring and/or ultrasonication). Saturation columns/systems or passive dosing methods (20) can be used for achieving a suitably concentrated stock solution. The preference is to use a co-solvent-free test system; however, different test chemicals will possess varied physicochemical properties that will likely require different approaches for preparation of chemical exposure water. All efforts should be made to avoid solvents or carriers because: (1) certain solvents themselves may result in toxicity and/or undesirable or unexpected responses, (2) testing chemicals above their water solubility (as can frequently occur through the use of solvents) can result in inaccurate determinations of effective concentrations, (3) the use of solvents in longer-term tests can result in a significant degree of “biofilming” associated with microbial activity which may impact environmental conditions as well as the ability to maintain exposure concentrations and (4) the absence of historical data that demonstrate that the solvent does not influence the outcome of the study, use of solvents requires a solvent control treatment which has significant animal welfare implications as additional animals are required to conduct the test. For difficult to test chemicals, a solvent may be employed as a last resort, and the OECD Guidance Document on Aquatic Toxicity Testing of Difficult Substances and Mixtures should be consulted (21) to determine the best method. The choice of solvent will be determined by the chemical properties of the test chemical and the availability of historical control data on the solvent. In the absence of historical data, the suitability of a solvent should be determined prior to conducting the definitive study. In the event that use of a solvent is unavoidable, and microbial activity (biofilming) occurs, recommend recording/reporting of the biofilming per tank (at least weekly) throughout the test. Ideally, the solvent concentration should be kept constant in the solvent control and all test treatments. If the concentration of solvent is not kept constant, the highest concentration of solvent in the test treatment should be used in the solvent control. In cases where a solvent

carrier is used, maximum solvent concentrations should not exceed 100 µl/l or 100 mg/l (21), and it is recommended to keep solvent concentration as low as possible (*e.g*, ≤ 20 µl/l) to avoid potential effects of the solvent on endpoints measured (22).

Test animals

Test species

22. The test species is *X. laevis* because this is: (1) routinely cultured in laboratories worldwide, (2) easily obtainable through commercial suppliers and (3) capable of having its genetic sex determined.

Adult care and breeding

23. Appropriate care and breeding of *X. laevis* is described by a standardised guideline (23). Housing and care of *X. laevis* are also described by Read (24). To induce breeding, three to five pairs of adult females and males are injected intraperitoneally with human chorionic gonadotropin (hCG). Female and male specimens are injected with e.g., approximately 800-1000 IU and 500-800 IU, respectively, of hCG dissolved in 0.6-0.9% saline solution (or frog Ringer's solution, an isotonic saline for use with amphibians; www.hermes.mbl.edu/biologicalbulletin/compendium/comp-RGR.html). Injection volumes should be about 10 µl/g body weight (~1000 µl). Afterwards, induced breeding pairs are held in large tanks, undisturbed and under static conditions to promote amplexus. The bottom of each breeding tank should have a false bottom of stainless steel mesh (e.g., 1.25 cm openings) which permits the eggs to fall to the bottom of the tank. Frogs injected with hCG in the late afternoon will usually deposit most of their eggs by mid-morning of the next day. After a sufficient quantity of eggs is released and fertilised, adults should be removed from the breeding tanks. Eggs are then collected and jelly coats are removed by L-cysteine treatment (23). A 2% L-cysteine solution should be prepared and pH adjusted to 8.1 with 1 M NaOH. This 21 ^{o}C solution is added to a 500 ml Erlenmeyer flask containing the eggs from a single spawn and swirled gently for one to two minutes and then rinsed thoroughly 6-8 times with 21 °C culture water. The eggs are then transferred to a crystallising dish and determined to be > 70% viable with minimal abnormalities in embryos exhibiting cell division.

TEST DESIGN

Test concentrations

24. It is recommended to use a minimum of four chemical concentrations and appropriate controls (including solvent controls, if necessary). Generally, a concentration separation (spacing factor) not exceeding 3.2 is recommended.

25. For the purposes of this test, results from existing amphibian studies should be used to the extent possible in determining the highest test concentration so as to avoid concentrations

that are overtly toxic. Information from, for example, quantitative structure-activity relationships, read across and data from existing amphibian studies such as the Amphibian Metamorphosis Assay, test method C.38 (25) and the Frog Embryo Teratogenesis Assay - *Xenopus* (23) and/or fish tests such as test methods C.48, C.41 and C.49 (26) (27) (28) may contribute toward setting this concentration. Prior to running the LAGDA a range finding experiment may be conducted. It is recommended that the range-finding exposure is initiated within 24 hours of fertilisation and continued for 7-14 days (or more, if needed), and the test concentrations are set such that the intervals between test concentrations are no greater than a factor of 10. The results of the range finding experiment should serve to set the highest test concentration in the LAGDA. Note that if a solvent has to be used, then the suitability of the solvent (i.e. whether it may have an impact on the outcome of the study) could be determined as part of the range finding study.

Replicates within treatment groups and controls

26. A minimum of four replicate tanks per test concentration and a minimum of eight replicates for the controls (and solvent control, if needed) should be used (i.e., the number of replicates in the control and any solvent control should be twice as large as the number of replicates of each treatment group, to ensure appropriate statistical power). Each replicate should contain no more than 20 animals. The minimum number of animals processed would be 15 (5 for NF stage 62 sub-sample and 10 juveniles). However, additional animals are added to each replicate to factor in the possibility for mortality while maintaining the critical number of 15.

PROCEDURE

Assay overview

27. The assay is initiated with newly spawned embryos (NF stage 8-10) and continues into juvenile development. Animals are examined daily for mortality and any sign of abnormal behaviour. At NF stage 62, a larval sub-sample (up to 5 animals per replicate) is collected and various endpoints are examined (Table 1). After all animals have reached NF stage 66, i.e. completion of metamorphosis (or after 70 days from the assay initiation, whichever comes first), a cull is carried out at random (but without sub-sampling) to reduce the number of animals (10 per tank) (see paragraph 43), and the remaining animals continue exposure until 10 weeks after the median time to NF stage 62 in the control. At test termination (juvenile sampling) additional measurements are made (Table 1).

Exposure conditions

28. A complete summary of test parameters can be found in Appendix 3. During the exposure period, dissolved oxygen, temperature, and pH of test solutions should be measured daily. Conductivity, alkalinity, and hardness are measured once a month. For the water temperature of test solutions, the inter-replicate and inter-treatment differentials (within

one day) should not exceed 1.0 °C. Also, for pH of test solutions, the inter-replicate and inter-treatment differentials should not exceed 0.5.

29. The exposure tanks may be siphoned on a daily basis to remove uneaten food and waste products, being careful to avoid cross-contamination of tanks. Care should be used to minimise stress and trauma to the animals, especially during movement, cleaning of aquaria, and manipulation. Stressful conditions/activities should be avoided such as loud and/or incessant noise, tapping on aquaria, vibrations in the tank.

Duration of exposure to the test chemical

30. The exposure is initiated with newly spawned embryos (NF stage 8-10) and continued until ten weeks after the median time to NF stage 62 (≤ 45 days from the assay initiation) in control group. Generally, the duration of the LAGDA is 16 weeks (maximum 17 weeks).

Initiation of assay

31. Parent animals used for the initiation of the assay should have previously been shown to produce offspring that can be genetically sexed (Appendix 5). After spawning of adults, embryos are collected, cysteine-treated to remove the jelly coat and screened for viability (23). Cysteine treatment allows the embryos to be handled during screening without sticking to surfaces. Screening takes place under a dissecting microscope using an appropriately sized eye dropper to remove non-viable embryos. It is preferred that a single spawn resulting in greater than 70% viability be used for the test. Embryos at NF stage 8-10 are randomly distributed into exposure treatment tanks containing an appropriate volume of dilution water until each tank contains 20 embryos. Embryos should be carefully handled during this transfer in order to minimise handling stress and to avoid any injury. At 96 hours post fertilisation, the tadpoles should have moved up the water column and begun clinging to the sides of the tank.

Feeding regime

32. Feed and feeding rate change during different life stages of *X. laevis* are a very important aspect of the LAGDA protocol. Excessive feeding during the larval phase typically results in increased incidences and severity of scoliosis (Appendix 8) and should be avoided. Conversely, inadequate feeding during the larval phase results in highly variable developmental rates among controls potentially compromising statistical power or confounding test results. Appendix 4 provides recommended larval and juvenile diet and feeding regimes for *X. laevis* in flow-through conditions, but alternatives are permissible providing the test organisms grow and develop satisfactorily. It is important to note that if endocrine-specific endpoints are being measured, feed should be free of endocrine-active substances such as soy meal.

Larval feeding

33. The recommended larval diet consists of trout starter feeds, *Spirulina* algae discs and goldfish crisps (*e.g.*, TetraFin® flakes, Tetra, Germany) blended together in culture (or dilution) water. This mixture is administered three times daily on weekdays and once daily on weekends. Tadpoles are also fed live brine shrimp, *Artemia* spp., 24-hour-old nauplii, twice daily on weekdays and once daily on the weekends starting on day 8 post-fertilisation. The larval feeding, which should be consistent in each test vessel, should allow appropriate growth and development for test animals in order to ensure reproducibility and transferability of the assay results: (1) the median time to NF stage 62 in controls should be ≤ 45 days and (2) a mean weight within 1.0 ± 0.2 g at NF stage 62 in controls is recommended.

Juvenile feeding

34. Once metamorphosis is complete, the feeding regime consists of premium sinking frog food, e.g., Sinking Frog Food -3/32 (Xenopus Express, FL, USA) (Appendix 4). For froglets (early juveniles), the pellets are briefly run in a coffee grinder, blender or crushed with a mortar and pestle in order to reduce their size. Once juveniles are large enough to consume full pellets, grinding or crushing is no longer necessary. The animals should be fed once per day. The juvenile feeding should allow appropriate growth and development of the organisms: a mean weight within 11.5 ± 3 g in control juveniles at the termination of the assay is recommended.

Analytical chemistry

35. Prior to initiation of the assay, the stability of the test chemical (*e.g.*, solubility, degradability, and volatility) and all analytical methods needed should be established *e.g.*, using existing information or knowledge. When dosing via the dilution water, it is recommended that test solutions from each replicate tank be analysed prior to test initiation to verify system performance. During the exposure period, the concentrations of the test chemical are determined at appropriate intervals, preferably every week for at least one replicate in each treatment group, rotating between replicates of the same treatment group every week. It is recommended that results be based on measured concentrations. However, if concentration of the test chemical in solution has been satisfactorily maintained within ± 20% of the nominal concentration throughout the test, then the results can either be based on nominal or measured values. Also, the coefficient of variation (CV) of the measured test concentrations over the entire test period within a treatment should be maintained at 20% or less in each concentration. When the measured concentrations do not remain within 80-120% of the nominal concentration (for example, when testing highly biodegradable or adsorptive chemicals), the effect concentrations should be determined and expressed relative to the arithmetic mean concentration for flow-through tests.

36. The flow rates of dilution water and stock solution should be checked at appropriate intervals (*e.g.* three times a week) throughout the exposure duration. In the case of

chemicals which cannot be detected at some or all of the nominal concentrations, (*e.g.*, due to rapid degradation or adsorption in the test vessels, or by marked chemical accumulation in the bodies of exposed animals), it is recommended that the renewal rate of the test solution in each chamber be adapted to maintain test concentrations as constant as possible.

Observations and endpoint measurements

37. The endpoints evaluated during the course of the exposure are those indicative of toxicity including mortality, abnormal behaviour such as clinical signs of disease and/or general toxicities, and growth determinations (length and weight), as well as pathology endpoints which may respond to both general toxicity and endocrine modes of action targeting oestrogen-, androgen-, or thyroid-mediated pathways. In addition, plasma VTG concentration may be optionally measured at the termination of the assay. Measurement of VTG may be useful in understanding study results in the context of endocrine mechanisms for suspected EDCs. The endpoints and timing of measurements are summarised in Table 1.

Table 1: Endpoint overview of the LAGDA

Endpoints*	**Daily**	**Interim Sampling (Larval sampling)**	**Test Termination (Juvenile sampling)**
Mortality and abnormalities	X		
Time to NF stage 62		X	
Histo(patho)logy (thyroid gland)		X	
Morphometrics (growth in weight and length)		X	X
Liver-somatic index (LSI)			X
Genetic/phenotypic sex ratios			X
Histopathology (gonads, reproductive ducts, kidney and liver)			X
Vitellogenin (VTG) (optional)			X

* All endpoints are analysed statistically.

Mortality and daily observations

38. All test tanks should be checked daily for dead animals and mortalities recorded for each tank. Dead animals should be removed from the test tank as soon as observed. The developmental stage of dead animals should be categorised as either pre-NF stage 58 (pre-forelimb emergence), NF stage 58-NF stage 62, NF stage 63-NF stage 66 (between NF stage 62 and complete tail absorption), or post-NF stage 66 (post-larval). Mortality rates exceeding 20% may indicate inappropriate test conditions or overtly toxic effects of the

test chemical. The animals tend to be most sensitive to non-chemical induced mortality events during the first few days of development after the spawning event and during metamorphic climax. Such mortality could be apparent from the control data.

39. In addition, any observation of abnormal behaviour, grossly visible malformations (*e.g.*, scoliosis), or lesions should be recorded. Observations of scoliosis should be counted (incidence) and graded with respect to severity (*e.g.*, not remarkable – NR, minimal – 1, moderate – 2, severe – 3; Appendix 8). Efforts should be made to ensure that the prevalence of moderate and severe scoliosis is limited (*e.g.*, below 10% in controls) throughout the study, although greater prevalence of control abnormalities would not necessarily be a reason for stopping the test. Normal behaviour for larval animals is characterised by suspension in the water column with tail elevated above the head, regular rhythmic tail fin beating, periodic surfacing, operculating, and being responsive to stimuli. Abnormal behaviours would include, for example, floating on the surface, lying on the bottom of the tank, inverted or irregular swimming, lack of surfacing activity, and being nonresponsive to stimuli. For post-metamorphic animals, in addition to the above abnormal behaviours, gross differences in food consumption between treatments should be recorded. Gross malformations and lesions could include morphological abnormalities (*e.g.*, limb deformities), haemorrhagic lesions, abdominal oedema, and bacterial or fungal infections, to name a few. The occurrences of lesions on the head of juveniles, just posterior to the nostrils, may be indications of insufficient humidity levels. These determinations are qualitative and should be considered akin to clinical signs of disease/stress and made in comparison to control animals. If the rate of occurrence is greater in exposed tanks than in the controls, then these should be considered as evidence for overt toxicity.

Larval sub-sampling

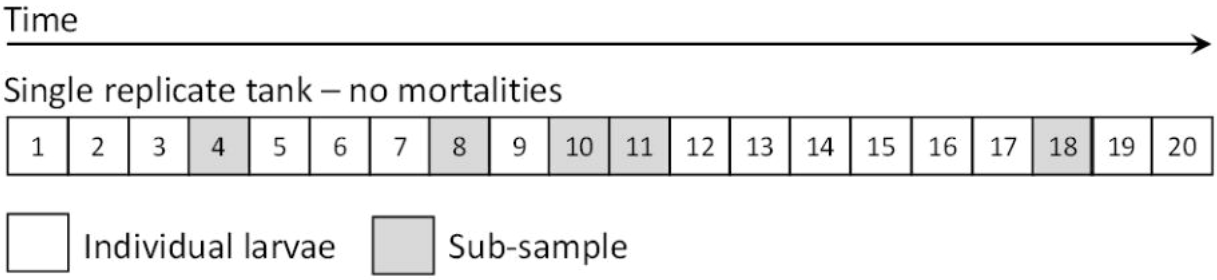
Outline of larval sub-sampling:

40. The tadpoles that have reached NF stage 62 should be removed from the tanks and either sampled or moved to the next part of the exposure in a new tank, or physically separated from the remaining tadpoles in the same tank with a divider. Tadpoles are checked daily, and the study day on which an individual tadpole reaches NF stage 62 is recorded. The defining characteristic for use in this assessment is the shape of the head. Once the head has become reduced in size such that it is visually approximately the same width as the trunk of the tadpole and forelimbs at the level of the middle of the heart, then that individual would be counted as having attained NF stage 62.

41. The goal is to sample a total of five NF stage 62 tadpoles per replicate tank. This should be performed entirely at random, but decided *a priori*. A hypothetical example of a replicate tank is provided in **Figure 1**. Should there be 20 surviving tadpoles in a particular tank when the first individual reaches NF stage 62, five random numbers should be chosen from 1-20. Tadpole #1 is the first individual to reach NF stage 62 and tadpole #20 is the last

individual in a tank to reach NF stage 62. Likewise, if there are 18 surviving larvae in a tank, five random numbers should be chosen from 1-18. This should be performed for every replicate tank when the first individual on test reaches NF stage 62. If there are mortalities during the NF stage 62 sampling, the remaining samples need to be re-randomised based on how many larvae are left <NF stage 62 and how many more samples are needed to reach a total of five samples from that replicate. On the day a tadpole reaches NF stage 62, reference to the prepared sampling chart is made to determine whether that individual is sampled or physically separated from the remaining tadpoles for continued exposure. In the example provided (Figure 1), the first individual to reach NF stage 62 (i.e. box #1) is physically separated from the other larvae, continues exposure and the study day on which that individual reached NF stage 62 is recorded. Subsequently, individuals #2 and #3 are treated the same way as #1 and then individual #4 is sampled for growth and thyroid histology (according to this example). This procedure continues until the 20th individual either joins the rest of the post-NF stage 62 individuals or is sampled. The random procedure used must give each organism on test equal probability of being selected. This can be achieved by using any randomising method, but also requires that each tadpole be netted at some point throughout the NF stage 62 sub-sampling period.

Figure 1: Hypothetical example of NF stage 62 sampling regime for a single replicate tank.



42. For the larval sub-sampling, the endpoints obtained are: (1) time to NF stage 62 (i.e., number of days between fertilisation and NF stage 62), (2) external abnormalities, (3) morphometrics (*e.g.*, weight and length) and (4) thyroid histology.

Humane killing of tadpoles

43. The sub-sample of NF stage 62 tadpoles (5 individuals per replicate) should be euthanised by immersion for 30 minutes in appropriate amounts (*e.g.*, 500 ml) of anaesthetic solution (*e.g.*, 0.3% solution of MS-222, tricaine methane sulfonate, CAS.886-86-2). MS-222 solution should be buffered with sodium bicarbonate to a pH of approximately 7.0 because unbuffered MS-222 solution is acidic and irritating to frog skin resulting in poor absorption and unnecessary additional stress to the organisms.

44. Using a mesh dip net, a tadpole is removed from the experimental chamber and transported (placed) into the euthanasia solution. The animal is properly euthanised and is ready for necropsy when it is unresponsive to external stimuli such as pinching the hind limb with a pair of forceps.

Morphometrics (weight and length)

45. Measurements of wet weight (nearest mg) and snout-to-vent length (SVL) (nearest 0.1 mm) for each tadpole should be made immediately after it becomes non-responsive by anaesthesia (Figure 2a). Image analysis software may be used to measure SVL from a photograph. Tadpoles should be blotted dry before weighing to remove excess adherent water. After measurements of body size (weight and SVL) are made, any gross morphological abnormalities and/or clinical signs of toxicity such as scoliosis (see Appendix 8), petechiae and haemorrhage should be recorded or noted, and digital documentation is recommended. Note that petechiae are small red or purple haemorrhages in skin capillaries.

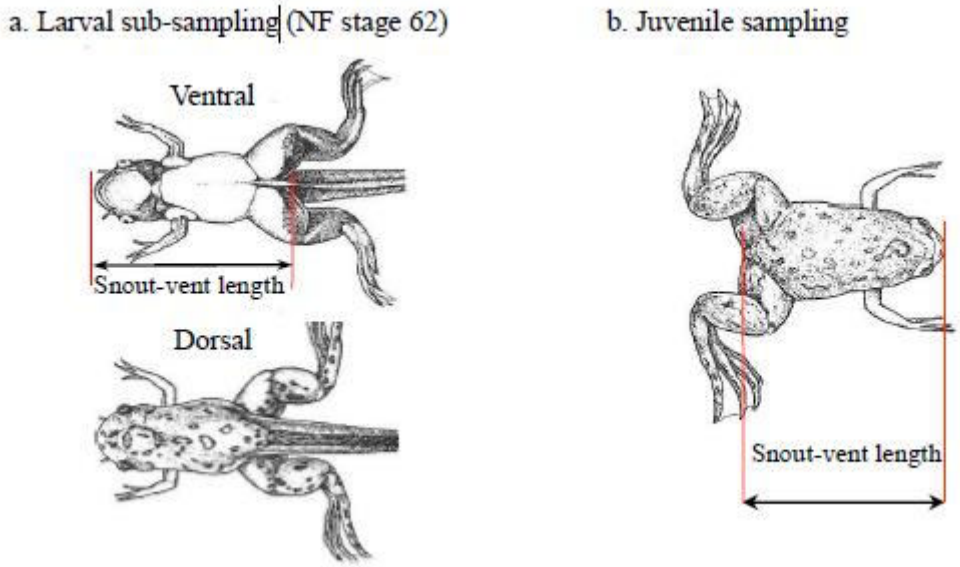
Tissue Collection and Fixation

46. For the larval sub-sample, thyroid glands are assessed for histology. The lower torso posterior to the forelimbs is removed and discarded. The trimmed carcass is fixed in Davidson's fixative. The volume of fixative in the container should be at least 10 times the approximate volume of the tissues. Appropriate agitation or circulation of the fixative should be achieved to adequately fix the tissues of interest. All tissues remain in Davidson's fixative for at least 48 hours, but no longer than 96 hours, at which time they are rinsed in deionised water and stored in 10% neutral buffered formalin (1) (29).

Thyroid histology

47. Each larval sub-sample (tissues fixed) is histologically assessed for thyroid glands, i.e., diagnosis and severity grading (29) (30).

Figure 2: Landmarks for measuring snout-vent length for the LAGDA in NF Stage 62 (a) and juvenile frogs (b). The defining characteristics of NF stage 62 (a): the head is the same width as the trunk, the olfactory nerve length is shorter than the diameter of the olfactory bulb (dorsal view), and the forelimbs are at the level of the heart (ventral view). Images adapted from Nieuwkoop and Faber (1994).



End of larval exposure

48. Given the initial number of tadpoles, it is expected that there will likely be a small percentage of individuals that do not develop normally and do not complete metamorphosis (NF stage 66) in a reasonable amount of time. The larval portion of the exposure should not exceed 70 days. Any tadpoles remaining at the end of this period should be euthanised (see para. 43), their wet weight and SVL measured, staged according to Nieuwkoop and Faber, 1994, and any developmental abnormalities noted.

Cull after NF stage 66

49. Ten individuals per tank should continue from NF stage 66 (complete tail resorption) until termination of the exposure. Therefore, after all animals have reached NF stage 66 or after 70 days (whichever occurs first), a cull should be conducted. Post NF stage 66 animals that will not continue the exposure should be selected at random.

50. Animals that are not selected for continued exposure are euthanised (see para. 43). Measurements of developmental stage, wet weight and SVL (Figure 2b) and a gross necropsy are conducted for each animal. The phenotypic sex (based on gonad morphology) is noted as female, male, or indeterminate.

Juvenile Sampling

Outline of juvenile sampling

51. The remaining animals continue exposure until 10 weeks after the median time to NF stage 62 in the dilution water (and/or solvent control if relevant) control. At the end of the exposure period, the remaining animals (maximum 10 frogs per replicate) are euthanised, and the various endpoints are measured or evaluated and recorded: (1) morphometrics (weight and length), (2) phenotypic/genotypic sex ratios, (3) liver weight (Liver-Somatic

Index), (4) histopathology (gonads, reproductive ducts, liver and kidney) and optionally (5) plasma VTG.

Humane killing of frogs

52. The juvenile samples, post-metamorphic frogs, are euthanised by an intraperitoneal injection of anaesthetic, *e.g.*, 10% MS-222 in an appropriate phosphate buffered solution. Frogs may be sampled after becoming unresponsive (usually around 2 min after injection, if 10% MS-222 is used in a dosage of 0.01 ml per g of frog). While the juvenile frogs could be immersed in a higher concentration of anaesthetic (MS-222), experience has shown that it takes longer for them to be anesthetised using this method and the duration may not be adequate to allow for sampling. Injection provides efficient, fast euthanasia prior to sampling. Sampling should not be started until lack of responsiveness of the frogs has been confirmed to ensure that the animals are dead. If frogs are showing signs of considerable suffering (very severe and death can be reliably predicted) and considered moribund, animals should be anaesthetised and euthanised and treated as mortality for data analysis. When a frog is euthanised due to morbidity, this should be noted and reported. Depending on when the frog is euthanised during the study, retaining the frog for histopathology analysis may be conducted (fixing the frog for possible histopathology).

Morphometrics (weight and length)

53. Measurements of wet weight and SVL (Figure 2b) are identical to those outlined for the larval sub-sampling.

Plasma VTG (option)

54. VTG is a widely accepted biomarker resulting from exposure to oestrogenic chemicals. For the LAGDA, plasma VTG optionally may be measured within juvenile samples (this may be particularly relevant if the test chemical is suspected of being an oestrogen).

55. The euthanised juvenile hind limbs are cut and blood is collected with a heparinised capillary (although alternative blood collection methods, such as cardiac puncture, may be suitable). The blood is expelled into a microcentrifuge tube (*e.g.*, 1.5 ml volume) and centrifuged to obtain plasma. The plasma samples should be stored at -70 °C or below until VTG determination. Plasma VTG concentration can be measured by an enzyme-linked immunosorbent assay (ELISA) method (Appendix 6), or by an alternative method such as mass spectrometry (31). Species specific antibodies are preferred due to greater sensitivity.

Genetic sex determination

56. The genetic sex of each juvenile frog is assessed based on the markers developed by Yoshimoto *et al.* (11). To determine the genetic sex, a portion (or whole) of one hind limb (or any other tissue) removed during dissection is collected and stored in a microcentrifuge tube (tissue samples from frogs can be obtained from any tissue). Tissue can be stored at -20°C or below until isolation of deoxyribose nucleic acid (DNA). The isolation of DNA

from tissues can be performed with commercially available kits and analysis for presence or absence of the marker is done by a polymerase chain reaction (PCR) method (Appendix 5). Generally, the concordance between histological sex and genotype across control animals at the juvenile sampling time point in control groups is more than 95%.

Tissue collection and fixation for histopathology

57. Gonads, reproductive ducts, kidneys and livers are collected for histological analysis during the final sampling. The abdominal cavity is opened, and the liver is dissected out and weighed. Next, the digestive organs (*e.g.*, stomach, intestines) are carefully removed from the lower abdomen to reveal the gonads, kidneys and reproductive ducts. Any gross morphological abnormalities in the gonads should be noted. Finally, the hind limbs should be removed if they have not previously been removed for blood collection. Collected livers and the carcass with the gonads left *in situ* should be immediately placed into Davidson's fixative. The volume of fixative in the container should be at least 10 times the approximate volume of the tissues. All tissues remain in Davidson's fixative for at least 48 hours, but no longer than 96 hours at which time they are rinsed in de-ionised water and stored in 10% neutral buffered formalin (1) (29).

Histopathology

58. Each juvenile sample is evaluated histologically for pathology in the gonads, reproductive ducts, kidneys and liver tissue, i.e., diagnosis and severity grading (32). The gonad phenotype is also derived from this evaluation (*e.g.*, ovary, testis, intersex), and together with individual genetic sex measurements, these observations can be used to calculate phenotypic/genotypic sex ratios.

DATA REPORTING

Statistical analysis

59. The LAGDA generates three forms of data to be statistically analysed: (1) quantitative continuous data (weight, SVL, LSI, VTG), (2) time-to-event data for developmental rates (i.e., days to NF stage 62 from assay initiation) and (3) ordinal data in the form of severity scores or developmental stages from histopathology evaluations.

60. It is recommended that the test design and selection of statistical test permit adequate power to detect changes of biological importance in endpoints where a NOEC or ECx is to be reported. Statistical analyses of the data (generally, replicate mean basis) should preferably follow procedures described in the document Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application (33). Appendix 7 of this test method provides the recommended statistical analysis decision tree and guidance for the treatment of data and in the choice of the most appropriate statistical test or model to use in the LAGDA.

61. The data from juvenile sampling (*e.g.*, growth, LSI) should be analysed for each genotypic sex separately since genotypic sex is determined for all frogs.

Data analysis considerations

Use of compromised replicates and treatments

62. Replicates and treatments may become compromised due to excess mortality from overt toxicity, disease, or technical error. If a treatment is compromised from disease or technical error, there should be three uncompromised treatments with three uncompromised replicates available for analysis. If overt toxicity occurs in the high treatment(s), it is preferable that at least three treatment levels with three uncompromised replicates are available for analysis (consistent with the Maximum Tolerated Concentration approach for OECD test guidelines (34)). In addition to mortality, signs of overt toxicity may include behavioural effects (e.g. floating on the surface, lying on the bottom of the tank, inverted or irregular swimming, lack of surfacing activity), morphological lesions (e.g. haemorrhagic lesions, abdominal oedema) or inhibition of normal feeding responses when compared qualitatively to control animals.

Solvent control

63. At the termination of the test, an evaluation of the potential effects of the solvent (if used) should be performed. This is done through a statistical comparison of the solvent control group and the dilution water control group. The most relevant endpoints for consideration in this analysis are growth determinants (weight and length), as these can be affected through generalised toxicities. If statistically significant differences are detected in these endpoints between the dilution water control and solvent control groups, best professional judgment should be used to determine if the validity of the test is compromised. If the two controls differ, the treatments exposed to the chemical should be compared to the solvent control unless it is known that comparison to the dilution water control is preferred. If there is no statistically significant difference between the two control groups it is recommended that the treatments exposed to the test chemical are compared with the pooled (solvent and dilution water control groups), unless it is known that comparison to either the dilution-water or solvent control group only is preferred.

Test report

64. The test report should include the following:

Test chemical:

- Physical nature and, where relevant, physicochemical properties;
- Mono-constituent substance:

  physical appearance, water solubility, and additional relevant physicochemical properties;

chemical identification, such as IUPAC or CAS name, CAS number, SMILES or InChI code, structural formula, purity, chemical identity of impurities as appropriate and practically feasible, etc. (including the organic carbon content, if appropriate).

- Multi-constituent substance, UVCBs and mixtures:

  characterised as far as possible by chemical identity (see above), quantitative occurrence and relevant physicochemical properties of the constituents.

Test species:

- Scientific name, strain if available, source and method of collection of the fertilised eggs and subsequent handling.
- Incidence of scoliosis in historical controls for the stock culture used.

Test conditions:

- Photoperiod(s);
- Test design (*e.g.*, chamber size, material and water volume, number of test chambers and replicates, number of test organisms per replicate);
- Method of preparation of stock solutions and frequency of renewal (the solubilising agent and its concentration should be given, when used);
- Method of dosing the test chemical (*e.g.*, pumps, diluting systems);
- The recovery efficiency of the method and the nominal test concentrations, the limit of quantification, the means of the measured values and their standard deviations in the test vessels and the method by which these were attained and evidence that the measurements refer to the concentrations of the test chemical in true solution;
- Dilution water characteristics: pH, hardness, temperature, dissolved oxygen concentration, residual chlorine levels (if measured), total iodine, total organic carbon (if measured), suspended solids (if measured), salinity of the test medium (if measured) and any other measurements made;
- The nominal test concentrations, the means of the measured values and their standard deviations;
- Water quality within test vessels, pH, temperature (daily) and dissolved oxygen concentration;
- Detailed information on feeding (*e.g.*, type of foods, source, amount given and frequency).

Results:

- Evidence that controls met the validity criteria;

- Data for the control (plus solvent control when used) and the treatment groups as follows: mortality and abnormality observed, time to NF stage 62, thyroid histology assessment (larval sample only), growth (weight and length), LSI (juvenile sample only), genetic/phenotypic sex ratios (juvenile sample only), histopathology assessment results for gonads, reproductive ducts, kidney and liver (juvenile sample only) and plasma VTG (juvenile sample only, if performed);
- Approach for the statistical analysis and treatment of data (statistical test or model used);
- No observed effect concentration (NOEC) for each response assessed;
- Lowest observed effect concentration (LOEC) for each response assessed (at $\alpha = 0.05$); ECx for each response assessed, if applicable, and confidence intervals (*e.g.*, 95%) and a graph of the fitted model used for its calculation, the slope of the concentration-response curve, the formula of the regression model, the estimated model parameters and their standard errors.
- Any deviation from the test method and deviations from the acceptance criteria, and considerations of potential consequences on the outcome of the test.

65. For the results of endpoint measurements, mean values and their standard deviations (on both replicate and concentration basis, if possible) should be presented.

66. Median time to NF stage 62 in controls should be calculated and presented as the mean of replicate medians and their standard deviation. Likewise, for treatments, a treatment median should be calculated and presented as the mean of replicate medians and their standard deviation.

LITERATURE

(1) U.S. Environmental Protection Agency (2013). Validation of the Larval Amphibian Growth and Development Assay: Integrated Summary Report.

(2) OECD (2012a). Guidance Document on Standardised Test Guidelines for Evaluating Endocrine Disrupters. Environment, Health and Safety Publications, Series on testing and assessment (No 150) Organisation for Economic Cooperation and Development, Paris.

(3) Nieuwkoop PD and Faber J. (1994). Normal Table of *Xenopus laevis* (Daudin). Garland Publishing, Inc, New York, NY, USA.

(4) Kloas W and Lutz I. (2006). Amphibians as Model to Study Endocrine Disrupters. Journal of Chromatography A 1130: 16-27.

(5) Chang C, Witschi E. (1956). Genic Control and Hormonal Reversal of Sex Differentiation in *Xenopus*. Journal of the Royal Society of Medicine 93: 140-144.

(6) Gallien L. (1953). Total Inversion of Sex in *Xenopus laevis* Daud, Following Treatment with Estradiol Benzoate Administered During Larval Stage. Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences 237: 1565.

(7) Villalpando I and Merchant-Larios H. (1990). Determination of the Sensitive Stages for Gonadal Sex-Reversal in *Xenopus Laevis* Tadpoles. International Journal of Developmental Biology 34: 281-285.

(8) Miyata S, Koike S and Kubo T. (1999). Hormonal Reversal and the Genetic Control of Sex Differentiation in *Xenopus*. Zoological Science 16: 335-340.

(9) Mikamo K and Witschi E. (1963). Functional Sex-Reversal in Genetic Females of *Xenopus laevis*, Induced by Implanted Testes. Genetics 48: 1411.

(10) Olmstead AW, Kosian PA, Korte JJ, Holcombe GW, Woodis K and Degitz SJ. (2009)a. Sex reversal of the Amphibian, *Xenopus tropicalis*, Following Larval Exposure to an Aromatase Inhibitor. Aquatic Toxicology 91: 143-150.

(11) Yoshimoto S, Okada E, Umemoto H, Tamura K, Uno Y, Nishida-Umehara C, Matsuda Y, Takamatsu N, Shiba T and Ito M. (2008). A W-linked DM-Domain Gene, DM-W, Participates in Primary Ovary Development in *Xenopus Laevis*. Proceedings of the National Academy of Sciences of the United States of America 105: 2469-2474.

(12) Olmstead AW, Korte JJ, Woodis KK, Bennett BA, Ostazeski S and Degitz SJ. (2009)b. Reproductive Maturation of the Tropical Clawed Frog: *Xenopus tropicalis*. General and Comparative Endocrinology 160: 117-123.

(13) Tobias ML, Tomasson J and Kelley DB. (1998). Attaining and Maintaining Strong Vocal Synapses in Female *Xenopus laevis*. Journal of Neurobiology 37: 441-448.

(14) Qin ZF, Qin XF, Yang L, Li HT, Zhao XR and Xu XB. (2007). Feminizing/Demasculinizing Effects of Polychlorinated Biphenyls on the Secondary Sexual Development of *Xenopus Laevis*. Aquatic Toxicology 84: 321-327.

(15) Porter KL, Olmstead AW, Kumsher DM, Dennis WE, Sprando RL, Holcombe GW, Korte JJ, Lindberg-Livingston A and Degitz SJ. (2011). Effects of 4-*Tert*-Octylphenol on *Xenopus Tropicalis* in a Long Term Exposure. Aquatic Toxicology 103: 159-169.

(16) ASTM. (2002). Standard Guide for Conducting Acute Toxicity Tests on Test Materials with Fishes, Macroinvertebrates, and Amphibians. ASTM E729-96, Philadelphia, PA, USA.

(17) Chapter C.4 of this Annex, Ready Biodegradability Test.

(18) Chapter C.29 of this Annex, Ready Biodegradability - CO2 in sealed vessels (Headspace Test).

(19) Kahl MD, Russom CL, DeFoe DL and Hammermeister DE (1999). Saturation Units for Use in Aquatic Bioassays. Chemosphere 39: 539-551.

(20) Adolfsson-Erici M, Åkerman G, Jahnke A, Mayer P, McLachlan MS (2012). A flow-through passive dosing system for continuously supplying aqueous solutions of hydrophobic chemicals to bioconcentration and aquatic toxicity tests. Chemosphere, 86(6): 593-9.

(21) OECD (2000). Guidance Document on Aquatic Toxicity Testing of Difficult Substances and Mixtures. Environment, Health and Safety Publications, Series on testing and assessment (No 23), Organisation for Economic Cooperation and Development, Paris.

(22) Hutchinson TH, Shillabeer N, Winter MJ and Pickford DB. (2006). Acute and Chronic Effects of Carrier Solvents in Aquatic Organisms: A Critical Review. Review. Aquatic Toxicology 76: 69–92.

(23) ASTM (2004). Standard Guide for Conducting the Frog Embryo Teratogenesis Assay - *Xenopus* (FETAX). ASTM E1439 - 98, Philadelphia, PA, USA.

(24) Read BT (2005). Guidance on the Housing and Care of the African Clawed Frog *Xenopus Laevis*. Royal Society for the Prevention of Cruelty to Animals (RSPCA), Horsham, Sussex, U.K., 84 pp.

(25) Chapter C.38 of this Annex, Amphibian Metamorphosis Assay.

(26) Chapter C.48 of this Annex, Fish Short Term Reproduction Assay.

(27) Chapter C.41 of this Annex, Fish Sexual Development Test.

(28) Chapter C.49 of this Annex, Fish Embryo Acute Toxicity (FET) Test.

(29) OECD (2007). Guidance Document on Amphibian Thyroid Histology.Environment, Health and Safety Publications, Series on Testing and Assessment. (No 82) Organisation for Economic Cooperation and Development, Paris.

(30) Grim KC, Wolfe M, Braunbeck T, Iguchi T, Ohta Y, Tooi O, Touart L, Wolf DC and Tietge J. (2009). Thyroid Histopathology Assessments for the Amphibian Metamorphosis Assay to Detect Thyroid-Active Substances, Toxicological Pathology 37: 415-424.

(31) Luna LG and Coady K.(2014). Identification of *X. laevis* Vitellogenin Peptide Biomarkers for Quantification by Liquid Chromatography Tandem Mass Spectrometry. Analytical and Bioanalytical Techniques 5(3): 194.

(32) OECD (2015). Guidance on histopathology techniques and evaluation. Environment, Health and Safety Publications, Series on Testing and Assessment (No 228), Organisation for Economic Cooperation and Development, Paris.

(33) OECD (2006). Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application. Environment, Health and Safety Publications, Series on testing and assessment (No 54), Organisation for Economic Cooperation and Development, Paris.

(34) Hutchinson TH, Bögi C, Winter MJ, Owens JW, 2009. Benefits of the Maximum Tolerated Dose (MTD) and Maximum Tolerated concentration (MTC) Concept in Aquatic Toxicology. Aquatic Toxicology 91(3): 197-202.

Appendix 1

DEFINITIONS

Apical endpoint: Causing effect at population level.

Chemical: A substance or a mixture

ELISA: Enzyme-Linked Immunosorbent Assay

ECx: (Effect concentration for x% effect) is the concentration that causes an x% of an effect on test organisms within a given exposure period when compared with a control. For example, an EC50 is a concentration estimated to cause an effect on a test end point in 50% of an exposed population over a defined exposure period.

dpf: Days post fertilization

Flow-through test: A test with continued flow of test solutions through the test system during the duration of exposure.

HPG axis: hypothalamic-pituitary-gonadal axis

IUPAC: International Union of Pure and Applied Chemistry.

Lowest observed effect concentration (LOEC) is the lowest tested concentration of a test chemical at which the chemical is observed to have a statistically significant effect (at $p < 0.05$) when compared with the control. However, all test concentrations above the LOEC should have a harmful effect equal to or greater than those observed at the LOEC. When these two conditions cannot be satisfied, a full explanation should be given for how the LOEC (and hence the NOEC) has been selected. Appendix 7 provides guidance.

Median Lethal Concentration (LC50): is the concentration of a test chemical that is estimated to be lethal to 50% of the test organisms within the test duration.

No observed effect concentration (NOEC) is the test concentration immediately below the LOEC, which when compared with the control, has no statistically significant effect ($p < 0.05$), within a stated exposure period.

SMILES: Simplified Molecular Input Line Entry Specification.

Test chemical: Any substance or mixture tested using this Test Method.

UVCB: Substances of unknown or variable composition, complex reaction products or biological materials.

VTG: Vitellogenin is a phospholipoglycoprotein precursor to egg yolk protein that normally occurs in sexually active females of all oviparous species.

Appendix 2

SOME CHEMICAL CHARACTERISTICS OF AN ACCEPTABLE DILUTION WATER

Substance	Limit concentration
Particulate matter	5 mg/l
Total organic carbon	2 mg/l
Un-ionised ammonia	1 µg/l
Residual chlorine	10 µg/l
Total organophosphorous pesticides	50 ng/l
Total organochlorine pesticides plus polychlorinated biphenyls	50 ng/l
Total organic chlorine	25 ng/l
Aluminium	1 µg/l
Arsenic	1 µg/l
Chromium	1 µg/l
Cobalt	1 µg/l
Copper	1 µg/l
Iron	1 µg/l
lead	1 µg/l
Nickel	1 µg/l
Zinc	1 µg/l
Cadmium	100 ng/l
Mercury	100 ng/l
Silver	100 ng/l

Appendix 3

TEST CONDITIONS FOR THE LAGDA

1. Test species	*Xenopus laevis*
2. Test type	Continuous flow-through,
3. Water temperature	The nominal temperature is 21 °C. The mean temperature over the duration of the test is 21 ± 1 ºC (the inter-replicate and the inter-treatment differentials should not exceed 1.0 ºC)
4. Illumination quality	Fluorescent bulbs (wide spectrum) 600-2000 lux (lumens/m^2) at the water surface
5. Photoperiod	12 h light:12 h dark
6. Test solution volume and test vessel (tank)	4-10 l (minimum 10–15 cm water depth) Glass or stainless steel tank
7. Volume exchanges of test solutions	Constant, in consideration of both the maintenance of biological conditions and chemical exposure (*e.g.*, 5 tank volume renewal per day)
8. Age of test organisms at initiation	Nieuwkoop and Faber (NF) stage 8-10
9. Number. of organisms per replicate	20 animals (embryos)/tank (replicate) at exposure initiation and 10 animals (juveniles)/tank (replicate) after NF stage 66 to exposure termination
10. Number of treatments	Minimum 4 test chemical treatments plus appropriate control(s)
11. Number of replicates per treatment	4 replicates per treatment for test chemical and 8 replicates for control(s)
12. Number of organisms per test concentration	Minimum 80 animals per treatment for test chemical and minimum 160 animals for control(s)
13. Dilution water	Any water that permits normal growth and development of *X. laevis* (*e.g.*, spring water or charcoal-filtered tap water)
14. Aeration	None required, but aeration of the tanks may be necessary if dissolved oxygen levels drop below recommended limits and increases in flow of test solution is maximised.
15. Dissolved oxygen of test solution	Dissolved oxygen: ≥ 40 % of air saturation value or ≥ 3.5 mg/l
16. pH of test solution	6.5-8.5 (the inter-replicate and the inter-treatment differentials should not exceed 0.5)

17. Hardness and alkalinity of test solution	10-250 mg $CaCO_3$/l
18. Feeding regime	(See Appendix 4)
19. Exposure period	From NF stage 8-10 to ten weeks after the median time to NF stage 62 in water and/or solvent control group (maximum 17 weeks)
20. Biological endpoints	Mortality (and abnormal appearances), time to NF stage 62 (larval sample), thyroid histology assessment (larval sample), growth (weight and length), liver-somatic index (juvenile sample), genetic/phenotypic sex ratios (juvenile sample), histopathology for gonads, reproductive ducts, kidney and liver (juvenile sample) and plasma vitellogenin (juvenile sample, optional)
21. Test validity criteria	Dissolved oxygen should be > 40% air saturation value; mean water temperature should be 21 ± 1 ºC and the inter-replicate and -treatment differentials should be < 1.0 ºC; pH of test solution should be ranged between 6.5 and 8.5; the mortality in control should be ≤ 20% in each replicate, and the mean time to NF stage 62 in control should be ≤ 45 days; the mean weight of test organisms at NF stage 62 and at the termination of the assay in controls and solvent controls (if used) should reach 1.0 ± 0.2 and 11.5 ± 3 g, respectively; evidence should be available to demonstrate that the concentrations of the test chemical in solution have been satisfactorily maintained within ± 20% of the mean measured values.

Appendix 4

FEEDING REGIME

It should be noted that although this feeding regime is recommended, alternatives are permissible providing the test organisms grow and develop at an appropriate rate.

Larval feeding

Preparation for larval diet

A. 1:1 (v/v) Trout Starter: algae/TetraFin® (or equivalent) ;
   1. Trout Starter: blend 50 g of Trout Starter (fine granules or powder) and 300 ml of suitable filtered water on a high blender setting for 20 seconds
   2. Algae/TetraFin® (or equivalent) mixture: blend 12 g spirulina algae disks and 500 ml filtered water on a high blender setting for 40 seconds, blend 12 g Tetrafin® (or equivalent) with 500 ml filtered water and then combine these to make up 1 L of 12 g/l spirulina algae and 12 g/l Tetrafin®(or equivalent)
   3. Combine equal volumes of the blended Trout Starter and the algae/TetraFin®(or equivalent) mixture

B. Brine shrimp:

15 ml brine shrimp eggs are hatched in 1 l of salt water (prepared by adding 20 ml of NaCl to 1 l deionised water). After aerating 24 hours at room temperature under constant light, the brine shrimp are harvested. Briefly, the brine shrimp are allowed to settle for 30 min by stopping aeration. Cysts that float to the top of the canister are poured off and discarded, and the shrimp are poured through the appropriate filters and brought up to 30 ml with filtered water.

Feeding Protocol

Table 1 provides a reference regarding the type and amount of feed used during the larval stages of the exposure. The animals should be fed three times per day Monday through Friday and once per day on the weekends.

Table 1: Feeding regime for *X. laevis* larvae in flow-through conditions

Time* **(Post Fertilisation)**	**Trout Starter: algae/TetraFin®(or equivalent)**		**Brine Shrimp**	
	Weekday (3 times per day)	**Weekend (once per day)**	**Weekday (twice per day)**	**Weekend (once per day)**
Days 4-14 (in Weeks 0-1)	0.33 ml	1.2 ml	0.5 ml (from Day 8 to 15) 1 ml (from Day 16)	0.5 ml (from Day 8 to 15) 1 ml (from Day 16)
Week 2	0.67 ml	2.4 ml		

Week 3	1.3 ml	4.0 ml	1 ml	1 ml
Week 4	1.5 ml	4.0 ml	1 ml	1 ml
Week 5	1.6 ml	4.4 ml	1 ml	1 ml
Week 6	1.6 ml	4.6 ml	1 ml	1 ml
Week 7	1.7 ml	4.6 ml	1 ml	1 ml
Weeks 8-10	1.7 ml	4.6 ml	1 ml	1 ml

* Day 0 is defined as the day hCG injection is done.

Larval to juvenile diet transition

As larvae complete metamorphosis, they transition to a juvenile diet formulation explained below. While this transition is taking place, the larval diet should be reduced as the juvenile feed increases. This can be accomplished by proportionally decreasing the larval feed while proportionally increasing the juvenile feed as each group of five tadpoles surpass NF stage 62 and approach completion of metamorphosis at NF stage 66.

Juvenile feeding

Juvenile diet

Once metamorphosis is complete (stage 66), the feeding regime changes to 3/32 inch premium sinking frog food alone (Xenopus Express™, FL, USA), or equivalent.

Preparation of crushed pellet for larval to juvenile transition

Sinking frog food pellets are briefly run in a coffee grinder, blender or mortar and pestle in order to reduce the size of the pellets by approximately 1/3. Processing too long results in powder and is discouraged.

Feeding protocol

Table 2 provides a reference regarding the type and amount of feed used during juvenile and adult life stages. The animals should be fed once per day. It should be noted that as animals metamorphose, they continue receiving a portion of the brine shrimp until > 95% of animals complete metamorphosis.

The animals should not be fed on the day of test termination so feed does not confound weight measurements.

Table 2: Feeding regime for *X. laevis* juveniles in flow-through conditions. It should be noted that unmetamorphosed animals, including those whose metamorphosis has been delayed by the chemical treatment, cannot eat uncrushed pellets.

Time (Weeks post-median metamorphosis date)	**Crushed pellet (mg per froglet)**	**Whole pellet (mg per froglet)**

As animals complete metamorphosis	25	0
Weeks 0-1	25	28
Weeks 2-3	0	110
Weeks 4-5	0	165
Weeks 6-9	0	220

* The first day of Week 0 is the median metamorphosis date in control animals.

Appendix 5

GENETIC SEX DETERMINATION (GENETIC SEXING)

The method of genetic sexing for *Xenopus laevis* is based on Yoshimoto *et al.*, 2008. Procedures in detail on the genotyping can be obtained from this publication, if needed. Alternative methods (e.g. high-throughput qPCR) may be used if considered suitable.

X. laevis primers

DM-W marker

Forward: 5'-CCACACCCAGCTCATGTAAAG-3'

Reverse: 5'-GGGCAGAGTCACATATACTG-3'

Positive Control

Forward: 5'-AACAGGAGCCCAATTCTGAG-3'

Reverse: 5'-AACTGCTTGACCTCTAATGC-3'

DNA purification

Purify DNA from muscle or skin tissue using e.g., Qiagen DNeasy Blood and Tissue Kit (cat # 69506) or similar product according to kit instructions. DNA can be eluted from the spin columns using less buffer to yield more concentrated samples if deemed necessary for PCR. Note that DNA is quite stable, so care should be taken to avoid cross-contamination that could lead to mischaracterisation of males as females, or vice versa.

PCR

A sample protocol using JumpStart™ *Taq* from Sigma is outlined in **Table 1**.

Table 1: Sample protocol using JumpStart™ *Taq* from Sigma

Master Mix	1x (µl)	[Final]
NFW	11	-
10X Buffer	2.0	-
$MgCl_2$ (25mM)	2.0	2.5 mM
dNTP's (10mM each)	0.4	200 µM
Marker for primer (8 µM)	0.8	0.3 µM

Marker rev primer (8 µM)	0.8	0.3 µM
Control for primer (8 µM)	0.8	0.3 µM
Control rev primer (8 µM)	0.8	0.3 µM
JumpStart™ Taq	0.4	0.05 units/µl
DNA template	1.0	~200 pg/µl

Note: When preparing Master Mixes, prepare extra to account for any loss that may occur while pipetting (example: 25x should be used for only 24 reactions).

Reaction:

Master Mix 19.0 µl

Template 1.0 µl

Total 20.0 µl

Thermocycler Profile:

Step 1. 94 ºC 1 min

Step 2. 94 ºC 30 sec

Step 3. 60 ºC 30 sec

Step 4. 72 ºC 1 min

Step 5. Go to step 2. 35 cycles

Step 6. 72 ºC 1 min

Step 7. 4 ºC hold

PCR products can be run immediately in a gel or stored at 4 ºC.

Agarose Gel Electrophoresis (3%)(sample protocol)

50X TAE

Tris 24.2 g

Glacial acetic acid 5.71 ml

Na_2 (EDTA)·$2H_2O$ 3.72 g

Add water to 100 ml

1X TAE

H_2O	392 ml
50X TAE	8 ml

3:1 Agarose

3 parts NuSieve™ GTG™ agarose

1 part Fisher agarose low electroendosmosis (EEO)

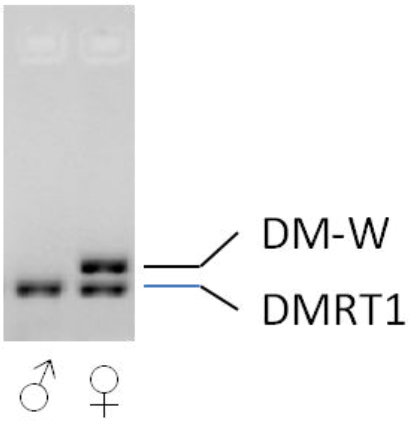
Method

1. Prepare a 3% gel by adding 1.2 g agarose mix to 43 ml 1X TAE. Swirl to disassociate large clumps.
2. Microwave agarose mixture until completely dissolved (avoid boiling over). Let cool slightly.
3. Add 1.0 µL ethidium bromide (10 mg/ml). Swirl flask. Note that ethidium bromide is mutagenic, so alternative chemicals should, in so far as is technically possible, be used for this step to minimise health risks to workers[1].
4. Pour gel into mould with comb. Cool completely.
5. Add gel to apparatus. Cover gel with 1X TAE.
6. Add 1 µl of 6x loading dye to each 10 µl PCR product.
7. Pipette samples into wells.
8. Run at 160 constant volts for ~20 minutes.

An agarose gel image showing the band patterns indicative of male and female individuals is shown in **Figure 1**.

Figure 1: Agarose gel image showing the band pattern indicative of a male (♂) individual (single band ~203 bp: DMRT1) and of a female (♀) individual (two bands at ~259 bp: DM-W and 203 bp:DMRT1).

[1] In accordance to Article 4.1 of Directive 2004/37/EC of the European Parliament and of the Council of 29 April 2004 on the protection of workers from the risks related to exposure to carcinogens or mutagens at work (Sixth individual Directive within the meaning of Article 16(1) of Council Directive 89/391/EEC) (OJ L 158, 30.4.2004, p. 50).



LITERATURE

Yoshimoto S, Okada E, Umemoto H, Tamura K, Uno Y, Nishida-Umehara C, Matsuda Y, Takamatsu N, Shiba T, Ito M. 2008. A W-linked DM-domain gene, DM-W, participates in primary ovary development in *Xenopus laevis*. Proceedings of the National Academy of Sciences of the United States of America 105: 2469-2474.

Appendix 6

MEASUREMENT OF VITELLOGENIN

The measurement of vitellogenin (VTG) is made using an enzyme-linked immunosorbent assay (ELISA) method which was originally developed for fathead minnow VTG (Parks *et al.*, 1999). Currently there are no commercially available antibodies for *X. laevis*. However, given the wealth of information for this protein and the availability of cost-effective commercial antibody production services, it is reasonable that laboratories can easily develop an ELISA to make this measure (Olmstead *et al.*, 2009). Also Olmstead *et al.* (2009) provide a description of the assay as modified for VTG in *X. tropicalis*, as shown below. The method uses an antibody made against *X. tropicalis* VTG, but it is known also to work for *X. laevis* VTG. It should be noted that non-competitive ELISAs can also be used, and that these may have lower detection limits than the method described below.

Materials and Reagents

- Preadsorbed 1st Antibody (Ab) serum

  Mix 1 part anti-*X. tropicalis* VTG 1st Ab serum with 2 parts control male plasma and leave at RT for ~ 75 minutes, put on ice for 30 min, centrifuge > 20K x G for 1 hour at 4 ºC, remove supernatant, aliquot, store at -20 ºC.

- 2nd Antibody

  Goat Anti-Rabbit IgG-HRP conjugate (e.g., Bio-Rad 172-1019)

- VTG Standard

  purified *X. laevis* VTG at 3.3 mg/ml.

- TMB (3,3',5,5' Tetramethyl-benzidine) (e.g., KPL 50-76-00; or Sigma T0440)
- Normal Goat Serum (NGS) (e.g., Chemicon® S26-100ml)
- 96 well EIA polystyrene microtiter plates (e.g., ICN: 76-381-04, Costar:53590, Fisher:07-200-35)
- 37 ºC hybridization oven (or fast equilibrating air incubator) for plates, water bath for tubes
- Other common laboratory equipment, chemicals, and supplies.

Recipes

Coating Buffer (50 mM Carbonate Buffer, pH 9.6):

$NaHCO_3$	1.26 g
Na_2CO_3	0.68 g
water	428 ml

10X PBS (0.1 M phosphate, 1.5 M NaCl):

$NaH_2PO_4 \cdot H_2O$ 0.83 g

$Na_2HPO_4 \cdot 7\ H_2O$ 20.1 g

NaCl 71 g

water 810 ml

Wash Buffer (PBST):

10X PBS 100 ml

water 900 ml

Adjust pH to 7.3 with 1 M HCl, then add 0.5 ml Tween-20

Assay Buffer:

Normal Goat Serum (NGS) 3.75 ml

Wash Buffer 146.25 ml

Sample collection

Blood is collected with a heparinised microhematocrit tube and placed on ice. After centrifugation for 3 minutes, the tube is scored, broken open, and the plasma expelled into 0.6 ml microcentrifuge tubes which contain 0.13 units of lyophilised aprotinin. (These tubes are prepared in advance by adding the appropriate amount of aprotinin, freezing, and lyophilising in a speed-vac at low heat until dry.) Store plasma at -80 °C until analysed.

Procedure for one plate

Coating the plate

Mix 20 µl of purified VTG with 22 ml of carbonate buffer (final 3 µg/ml). Add 200 µl to each well of a 96-well plate. Cover the plate with adhesive sealing film and allow to incubate at 37 ºC for 2 hours (or 4 ºC overnight).

Blocking the plate

Blocking solution is prepared by adding 2 ml of Normal Goat Serum (NGS) to 38 ml of carbonate buffer. Remove coating solution and shake dry. Add 350 µl of the blocking solution to each well. Cover with adhesive sealing film and incubate at 37 ºC for 2 hours (or at 4 °C overnight).

Preparation of standards

5.8 µl of purified VTG standard is mixed with 1.5 ml of assay buffer in a 12 x 75 mm borosilicate disposable glass test tube. This yields 12 760 ng/ml. Then a serial dilution is performed by adding 750 µl of the previous dilution to 750 µl of assay buffer to yield final concentrations of 12 760, 6380, 3190, 1595, 798, 399, 199, 100, and 50 ng/ml.

Preparation of Samples

Start with a 1:300 (e.g., combine 1 µl plasma with 299 µl of assay buffer) or 1:30 dilution of plasma into assay buffer. If a large amount of VTG is expected, additional or greater dilutions may be needed. Try to keep B/B_o within the range of standards. For samples without appreciable VTG, e.g., control males and females (which are all immature), use the 1:30 dilution. Samples diluted less than this may show unwanted matrix effects.

Additionally, it is recommended to run a positive control sample on each plate. This comes from a pool of plasma containing high induced levels of VTG. The pool is initially diluted in NGS, divided in aliquots and stored at -80 C. For each plate, an aliquot is thawed, diluted further in assay buffer and run similar to a test sample.

Incubation with 1st antibody

The 1st Ab is prepared by making a 1:2000 dilution of preadsorbed 1st Ab serum in assay buffer (e.g., 8 µl to 16 ml of assay buffer). Combine 300 µl of 1st Ab solution with 300 µl of sample/standard in a glass tube. The B_o tube is prepared similarly with 300 µl of assay buffer and 300 µl of antibody. Also, a NSB tube should be prepared using 600 µl of assay buffer only, i.e., no Ab. Cover the tubes with Parafilm and vortex gently to mix. Incubate in a 37 ºC water bath for 1 hour.

Washing the plate

Just before the 1st Ab incubation is complete, wash the plate. This is done by shaking out the contents and patting dry on absorbent paper. Then fill wells with 350 µl of wash solution, dump out, and pat dry. A multi-channel repeater pipette or plate washer is useful here. The wash step is repeated two more times for a total of three washes.

Loading the plate

After the plate has been washed, remove the tubes from the water bath and vortex lightly. Add 200 µl from each sample, standard, B_o, and NSB tube to duplicate wells of the plate. Cover plate with adhesive sealing film and allow to incubate for 1 hour at 37 ºC.

Incubation with the 2nd antibody

At the end of the incubation from the previous step, the plate should be washed three times again, like above. The diluted 2nd Ab is prepared by mixing 2.5 µl of 2nd Ab with 50 ml of assay buffer. Add 200 µl of diluted 2nd Ab to each well, seal like above, and incubate for 1 hour at 37 ºC.

Addition of substrate

After the incubation with the 2nd Ab is complete, wash the plate three times as described earlier. Then add 100 µl of TMB substrate to each well. Allow the reaction to proceed for 10 minutes, preferably out of bright light. Stop the reaction by adding 100 µl of 1 M phosphoric acid. This will change the colour from blue to an intense yellow. Measure the absorbance at 450 nm using a plate reader.

Calculate B/Bo

Subtract the average NSB value from all measurements. The B/B_o for each sample and standard is calculated by dividing the absorbance value (B) by the average absorbance of the B_o sample.

Obtain the standard curve and determine unknown amounts

Generate a standard curve with the aid of some computer graphing software (e.g., Slidewrite™ or Sigma Plot®) that will extrapolate quantity from B/B_o of sample based on B/B_o of standards. Typically, the amount is plotted on a log scale and the curve has a sigmoid shape. However, it may appear linear when using a narrow range of standards. Correct sample amounts for dilution factor and report as mg VTG/ml of plasma.

Determination of minimum detection limits (MDL)

Often, particularly in normal males, it will not be clear how to report results from low values. In these cases, the 95% "Confidence limits" should be used to determine if the value should be reported as zero or as some other number. If the sample result is within the confidence interval of the zero standard (B_o), the result should be reported as zero. The minimum detection level will be the lowest standard which is consistently different from the zero standard; that is, the two confidence intervals don't overlap. For any sample result which is within the confidence limit of the minimum detection level, or above, the calculated value will be reported. If a sample falls between the zero standard and the minimum detection level intervals, one half of the minimum detection level should be reported for the value of that sample.

LITERATURE

Olmstead AW, Korte JJ, Woodis KK, Bennett BA, Ostazeski S, Degitz SJ. 2009. Reproductive maturation of the tropical clawed frog: *Xenopus tropicalis*. General and Comparative Endocrinology 160: 117-123.

Parks LG, Cheek AO, Denslow ND, Heppell SA, McLachlan JA, LeBlanc GA, Sullivan CV. 1999. Fathead minnow (*Pimephales promelas*) vitellogenin: purification, characterisation and quantitative immunoassay for the detection of estrogenic compounds. Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology 123: 113-125.

Appendix 7

STATISTICAL ANALYSIS

The LAGDA generates three forms of data to be statistically analysed: (1) Quantitative continuous data, (2) Time-to-event data for developmental rates (Time to NF stage 62) and (3) Ordinal data in the form of severity scores or developmental stages from histopathology evaluations. The recommended statistical analysis decision tree for the LAGDA is shown in Figure 1. Also, some annotations which might be needed to conduct statistical analysis for the measurements from the LAGDA are indicated below. For the analysis decision tree, the results of measurements for mortality, growth (weight and length) and liver-somatic-index (LSI) should be analysed according to the "Other endpoints" branch.

Continuous data

Data for continuous endpoints should first be checked for monotonicity by rank transforming the data, fitting to an ANOVA model and comparing linear and quadratic contrasts. If the data are monotonic, a step-down Jonckheere-Terpstra trend test should be performed on replicate medians and no subsequent analyses should be applied. An alternative for data that are normally distributed with homogeneous variances is the step-down Williams' test. If the data are non-monotonic (quadratic contrast is significant and linear is not significant), they should be analysed using a mixed effects ANOVA model. The data should then be assessed for normality (preferably using the Shapiro-Wilk or Anderson-Darling test) and variance homogeneity (preferably using Levene's test). Both tests are performed on the residuals from the mixed effects ANOVA model. Expert judgment can be used in lieu of these formal tests for normality and variance homogeneity, though formal tests are preferred. If the data are normally distributed with homogeneous variance, then the assumptions of a mixed effect ANOVA are met and a significant treatment effect is determined from Dunnett's test. Where non-normality or variance heterogeneity is found, then the assumptions of Dunnett's test are violated and a normalising, variance stabilising transform is sought. If no such transform is found, then a significant treatment effect is determined with a Dunn's test. Whenever possible, a one-tailed test should be performed as opposed to a two-tailed test, but it requires expert judgment to determine which is appropriate for a given endpoint.

Mortality

Mortality data should be analysed for the time period encompassing the full test and should be expressed as proportion that died in any particular tank. Tadpoles that do not complete metamorphosis in the given time frame, those tadpoles that are in the larval sub-sample cohort, those juvenile frogs that are culled, and any animal that dies due to experimenter error should be treated as censored data and not included in the denominator of the percent calculation. Prior to any statistical analyses, mortality proportions should be arcsin-square root transformed. An alternative is to use the step-down Cochran-Armitage test, possibly with a Rao-Scott adjustment in the presence of overdispersion.

Weight and length (growth data)

Males and females are not sexually-dimorphic during metamorphosis so larval sub-sampling growth data should be analysed independent of gender. However, juvenile growth data should be analysed separately based on genetic sex. A log-transformation may be needed for these endpoints since log-normality of size data is not uncommon.

Liver-somatic-index (LSI)

Liver weights should be normalised as proportions of whole body weights (i.e., LSI) and analysed separately based on genetic sex.

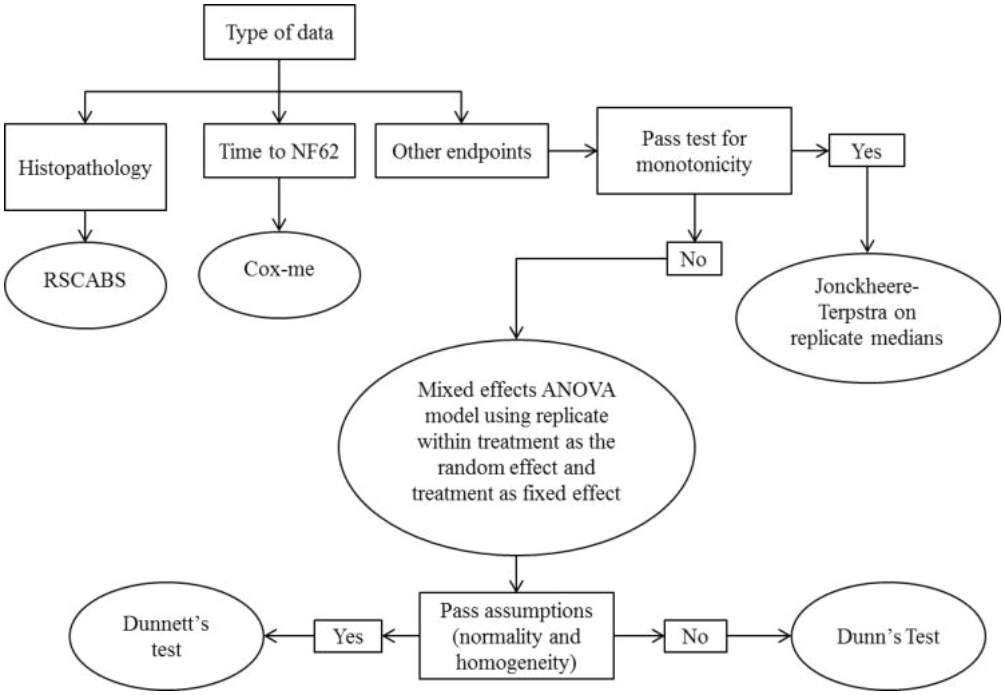
Time to NF stage 62

Time to metamorphosis data should be treated as time-to-event data, with any mortalities or individuals not reaching NF stage 62 in 70 days treated as right-censored data (i.e. the true value is greater than 70 days but the study ends before the animals had reached NF stage 62 in 70 days). Median time to NF stage 62 completion of metamorphosis in dilution water controls should be used to determine the test termination date. Median time to completion of metamorphosis could be determined by Kaplan-Meier product-limit estimators. This endpoint should be analysed using a mixed-effects Cox proportional hazard model that takes account of the replicate structure of the study.

Histopathology data (severity scores and developmental stages)

Histopathology data are in the form of severity scores or developmental stages. A test termed RSCABS (Rao-Scott Cochran-Armitage by Slices) uses a step-down Rao-Scott adjusted Cochran-Armitage trend test on each level of severity in a histopathology response (Green *et al.*, 2014). The Rao-Scott adjustment incorporates the replicate vessel experimental design into the test. The "by Slices" procedure incorporates the biological expectation that severity of effect tends to increase with increasing doses or concentrations, while retaining the individual subject scores and revealing the severity of any effect found. The RSCABS procedure not only determines which treatments are statistically different from controls (i.e., have more severe pathology than controls), but it also determines at which severity score the difference occurs thereby providing much needed context to the analysis. In the case of developmental staging of gonads and reproductive ducts, an additional manipulation should be applied to the data since an assumption of RSCABS is that severity of effect increases with dose. The effect observed could be a delay or acceleration of development. Therefore, developmental staging data should be analysed as reported to detect acceleration in development and then manually inverted prior to a second analysis to detect a delay in development.

Figure 1: Statistical analysis decision tree for LAGDA data.



LITERATURE

Green JW, Springer TA, Saulnier AN, Swintek J. 2014. Statistical analysis of histopathology endpoints. Environmental Toxicology and Chemistry 33, 1108-1116.

Appendix 8

CONSIDERATIONS FOR TRACKING AND MINIMISING THE OCCURRENCE OF SCOLIOSIS

Idiopathic scoliosis, usually manifesting as "bent tail" in *Xenopus laevis* tadpoles, may complicate morphological and behavioural observations in test populations. Efforts should be made to minimise or eliminate the incidence of scoliosis, both in stock and under test conditions. In the definitive test, it is recommended that the prevalence of moderate and severe scoliosis be less than 10%, to improve confidence that the test can detect treatment-related developmental effects in otherwise healthy amphibian larvae.

Daily observations during the definitive test should record both the incidence (individual count) and severity of scoliosis, when present. The nature of the abnormality should be described with respect to location (*e.g.*, anterior or posterior to the vent) and direction of curvature (*e.g.*, lateral or dorsal-to-ventral). Severity may be graded as follows:

(NR) Not remarkable: no curvature present

(1) Minimal: slight, lateral curvature posterior to the vent; apparent only at rest

(2) Moderate: lateral curvature posterior to the vent; visible at all times but does not inhibit movement

(3) Severe: lateral curvature anterior to the vent; OR any curvature that inhibits movement; OR any dorsal-to-ventral curvature

A US EPA FIFRA Scientific Advisory Panel (FIFRA SAP 2013) reviewed summary data for scoliosis in fifteen Amphibian Metamorphosis Assays with *X. laevis* (NF stage 51 through 60+) and provided general recommendations for reducing the prevalence of this abnormality in test populations. The recommendations are relevant to the LAGDA even though this test encompasses a longer developmental timeline.

Historical Spawning Performance

Generally, high quality, healthy adults should be used as breeding pairs; eliminating breeding pairs that produce offspring with scoliosis may minimise its occurrence over time. Specifically, minimising the use of wild-caught breeding stock may be beneficial. The LAGDA exposure period begins with NF stage 8-to-10 embryos, and it is not feasible to determine at the test outset whether given individuals will exhibit scoliosis. Thus, in addition to tracking the incidence of scoliosis in animals that are placed on test, historical clutch performance (including the prevalence of scoliosis in any larvae allowed to develop) should be documented. It may be useful to further monitor the portion of each clutch not used in a given study and to report these observations (FIFRA SAP 2013).

Water Quality

It is important to ensure adequate water quality, both in laboratory stock and during the test. In addition to water quality criteria routinely evaluated for aquatic toxicity tests, it may be useful to monitor for and to correct any nutrient deficiencies (*e.g.*, deficiency of vitamin C, calcium, phosphorus) or excess levels of selenium and copper, which are reported to cause scoliosis to varying degrees in laboratory-reared *Rana* sp. and *Xenopus* sp. (Marshall *et al.* 1980; Leibovitz *et al.* 1992; Martinez *et al.* 1992; as reported in FIFRA SAP 2013). The use of an appropriate dietary regimen (see Appendix 4), and regular tank cleaning, will generally improve water quality and health of the test specimens.

Diet

Specific recommendations for a dietary regimen, found to be successful in the LAGDA, are detailed in Appendix 4. It is recommended that feed sources be screened for biological toxins, herbicides, and other pesticides which are known to cause scoliosis in *X. laevis* or other aquatic animals (Schlenk and Jenkins 2013). For example, exposure to certain cholinesterase inhibitors has been associated with scoliosis in fish (Schultz *et al.* 1985) and frogs (Bacchetta *et al.* 2008).

LITERATURE

Bacchetta, R., P. Mantecca, M. Andrioletti, C. Vismara, and G. Vailati. 2008. Axial-skeletal defects caused by carbaryl in *Xenopus laevis* embryos. Science of the Total Environment 392: 110 – 118.

Schultz, T.W., J.N. Dumont, and R.G. Epler. 1985. The embryotoxic and osteolathyrogenic effects of semicarbazide. Toxicology 36: 185-198.

Leibovitz, H.E., D.D. Culley, and J.P. Geaghan. 1982. Effects of vitamin C and sodium benzoate on survival, growth and skeletal deformities of intensively culture bullfrog larvae (*Rana catesbeiana*) reared at two pH levels. Journal of the World Aquaculture Society 13: 322-328.

Marshall, G.A., R.L. Amborski, and D.D. Culley. 1980. Calcium and pH requirements in the culture of bullfrog (*Rana catesbeiana*) larvae. Journal of the World Aquaculture Society 11: 445-453.

Martinez, I., R. Alvarez, I. Herraez, and P. Herraez. 1992. Skeletal malformations in hatchery reared *Rana perezi* tadpoles. Anatomical Records 233(2): 314-320.

Schlenk, D., and Jenkins, F. 2013. Endocrine Disruptor Screening Prog (EDSP) Tier 1 Screening Assays and Battery Performance. US EPA FIFRA SAP Minutes No. 2013-03. May 21-23, 2013. Washington, DC. "