Krebs et al.:

Template for the Description of Cell-based Toxicological Test Methods to Allow Evaluation and Regulatory Use of the Data

Full article: *ALTEX 36*, 682-699. doi:10.14573/altex.1909271

Supplementary Data

Template for the description of cell-based toxicological test methods (ToxTemp)

S1.1 Introduction / preamble

Reproducible toxicology research necessitates the comprehensive description of the testing process. The necessary information can be grouped as belonging to (i) the overall test method description, (ii) the technical test procedure (as outlined in a standard operating procedure (SOP)), (iii) the characterization of test and reference materials/chemical and (iv) all issues relating to data processing and storage.

Here we focus mainly on the **test method description**. In toxicology, the term "test method" is used to describe a procedure used to obtain information on the potentially hazardous effects of a substance. A toxicological test method consists of four major components (i.e., (1) test system, (2) endpoint, (3) exposure scheme, (4) prediction model), and it produces a test result (information regarding the ability of a substance or agent to produce a specified biological/toxicological effect under specified conditions). A fifth component (5) of the test method gives the frame and boundaries of the method (it relates to the test purpose and applicability) and is also included here.

Even though most parts of the toxicological test method description template (ToxTemp) are generally valid, the details are dedicated to (cell-based) *in vitro* test methods used for the assessment of specific toxicological endpoints. In this context, it is important to refer especially to the Good *In Vitro* Method Practices (OECD, 2018). In this document, chapter 9 refers to reporting, with the key message: Good reporting of *in vitro* methods can only be achieved when all important details are recorded in a way that allows others to reproduce the work or reconstruct fully the *in vitro* method study. Particularly noteworthy is also chapter 5 (on test systems) with the key message that only few cell and tissue culture-based test systems have been used in regulatory approved test guideline methods due to reliability issues. Accordingly, GIVIMP devotes an entire chapter (chapter 8) to the performance of the test method and respective acceptance criteria. It is stated that *in vitro* method developers need to ensure that *in vitro* methods they design will produce good quality data, i.e., fit for purpose, thanks to a stringent assessment of the performance of the method.

Many of the items discussed here have already been addressed in other documents, e.g., in the DB-ALM methods summary (adapted from GD211) or the EURL-ECVAM test submission template (used for structuring information for test validation). Several earlier EU-funded projects devoted considerable resources to harmonize test method descriptions (Kinsner-Ovaskainen et al., 2009; Rovida et al., 2014), and similar activities are taking place in the USA (Flood et al., 2017). The overall reporting of *in vitro* experiments (i.e. data and methods) has been addressed by an NC3Rs initiative (Prior et al., 2019), by a large stakeholder workshop organized by CAAT-Europe (Hartung et al., 2019), and by the OECD (OECD harmonized templates for data reporting (OHT). OHT 201 on "Intermediate effects", published in August 2016 is especially relevant.[1]

Following the above ideas, this template was developed and reviewed by a large group of experts from the academic, regulatory, validation, industrial, and other relevant environments. The objective is to provide an easy to use, additional tool that facilitates high quality documentation of the status of (cell-based) *in vitro* test methods and their elements (e.g. test system). Moreover, the ToxTemp is intended to provide guidance for considerations around the use of test methods (or data therefrom) beyond the developer lab. For instance, the template contains information boxes important for test method transfer between laboratories or for scoring of the readiness of a test method for a specified application (Bal-Price et al., 2018). **In this sense, it targets broadly all those that expect their data to be used for important decisions (public health, commercial, regulatory, etc.).**

The foremost objective is to provide all stakeholders/test developers with a tool to self-evaluate where they stand. This distinguishes this document from the guidance documents cited above. The motivation to assemble the ToxTemp template comes from the observation that many researchers (independent of training, experience and position/affiliation) lack essential knowledge on test method descriptions. Moreover, they often have problems understanding and using the officially available guidance. In European projects, such as Acutetox, ESNATS, Reprotect, or EU-ToxRisk we found that the compliance with rules for high quality test method documentation is greatly improved when additional instructions are provided that facilitate information structuring and give clear explanations and examples of what needs to be provided.

doi:10.14573/altex.1909271s

[1] http://www.oecd.org/ehs/templates/harmonised-templates-intermediate-effects.htm

The information collection according to the ToxTemp should be considered complementary (Leist and Hengstler, 2018) to a standard operating procedure (SOP) elaborating technical details for test execution. The respective SOPs should be deposited in a high-quality database such as DB-ALM (EURL ECVAM Database Service on Alternative Methods to Animal Experimentation available at https://ecvamdbalm.jrc.ec.europa.eu/ (Roi, 2006). A potential outlook to future developments is the generation of an extended database containing method descriptions. This was initiated during the EU-ToxRisk project (with about 20 methods already recorded), but may be put on a broader basis. Another outlook is the development of a better understanding for requirements of test method documentation in academia and industry. Adoption of the principles laid out in ToxTemp in the scientific culture may improve overall research on *in vitro* methods/NAMs, and it may be envisaged that higher standards are being set also by journals for the submission of research articles or, e.g., for research projects financed by public financial resources (similar to the effects of the ARRIVE guidelines (Leung et al., 2018) in the field of *in vivo* experimentation and testing).

The information contained in the ToxTemp fulfils the requirements of OECD GD211 (OECD, 2017) for the description of non-guideline *in vitro* test methods (Hartung et al., 2019). It also contains additional information to not only satisfy regulatory requirements but also the practical implementation, transfer, and readiness assessment of the test methods (Bal-Price et al., 2018). Important terms and definitions can be found in Schmidt et al. (2017), Leist et al. (2010, 2012), or in the appendix of the present document. These definitions explain e.g. the difference between a “test method” and a “test system”.

Note that ToxTemp was designed to include information for various assay purposes. In some situations (early stages of test method development), some of the information is not available (or only partially available). Moreover, the ToxTemp is complementary to other documents (e.g. the test SOP), and respective fields may not be filled with extensive information (provided the info is available from the SOP), but only to provide a general overview.

Fields that are not mandatory, or that have only relative/partial requirements are labelled with an asterisk (*).

S1.2 ToxTemp

1. Overview

1.1 Descriptive full-text title

Provide a descriptive title using normal language without technical terms or acronyms.

Example: “Assay to test compound-derived impairment in neurite outgrowth in human mature dopaminergic neurons (NeuriTox; UKN4).”

1.2 Abstract

Please describe in no more than 200 words the following:

Which toxicological target (organ, tissue, physiological/biochemical function, etc.) is modelled? (8.1)

Which test system and readout(s) are used? (4.1; 5.2)

Which biological process(es) (e.g. neurite outgrowth, differentiation) and/or toxicological events (e.g. oxidative stress, cell death) are modelled/reflected by your test method? (8.1)

To which (human) adverse outcome(s) is your test method related or could be related? (8.1; 9.2; 9.3)

Which hazard(s) do(es) your test method (potentially) predict? (8.1; 8.6)

Does the test method capture an endpoint of current regulatory studies? (9.5)

If the method has undergone some form of validation/evaluation, give its status. (9.4)

Note: This section should give an overview. Details can be found in the respective chapters, as indicated by numbers in brackets.

***Important note**: Look at the Appendix before filling in the test method description form.*

Note: A test consists of several elements. If one is changed (e.g. the endpoint, duration of compound exposure, etc.), it requires a new submission with a different database name.

For the description here, choose one strictly-defined protocol / application, and refer all information to this setup. Possible variations can be indicated in chapter 6.8. If a variation (e.g. altered incubation time or other endpoint) is implemented, then such a variant must be deposited as a new test.

If a section / question is not applicable for your test method, state “not applicable”. If there is no information, please state “no information available”.

2. General information

2.1 Name of test method

Provide the original/published name, as well as the potential tradename.

2.2 Version number and date of deposition

Provide the original deposition date of first version and date of current version.

2.3 Summary of introduced changes in comparison to previous version(s)

This only applies to updated versions. If this is the original version, state "original version".

2.4 Assigned data base name

Normal text names often do not uniquely define the method. Therefore, each method should be assigned a clearly and uniquely defined data base name.

These are some example data base names generated in the EU-ToxRisk project:

UKN1a_DART_NPC_Diff_6D_02

UKN1b_DART_NPC_Diff_4D_01

UKN2a_DART_NC_Migr_24h_04

The name is assembled (in more generic terms) from the following elements:

Axa_B_C_D_E

Axa: mandatory part of the identifier allowing unambiguous identification

A: Abbreviation/acronym of the partner depositing the assay

x: Consecutive number (referring to the partner's assay number)

a: Sub-specifier (for variants, i.e. very similar assays but e.g. different readout or medium); not mandatory, but 'Axa' must be specific (i.e. clearly identifying) for each assay variant.

B: Indication of the main intended use (max. 5 letters), e.g. DART, Neuro, Liver, Lung, Renal, Redox, Stress...

C: Specifier of test system, e.g. cell type such as NPC (neural precursor cells), NC (neural crest), Hep (liver cells), REN (kidney cells), PUL (lung cells) (max. 4 letters)

D: Identification of test endpoint, e.g. Diff_6D = Differentiation for 6 days; exp_24 h = exposure for 24 hours; RNA_6h = transcriptome after 6 hours (use max. 15 signs altogether; if desired in 2-3 blocks), name (and acronym) of the project partner home organisation.

E: version number.

2.5 Name and acronym of the test depositor

Include affiliation.

2.6 Name and email of contact person

Provide the details of the principal contact person.

2.7 Name of further persons involved

For example, the principal investigator (PI) of the lab, the person who conducted the experiments, etc.

2.8 Reference to additional files of relevance

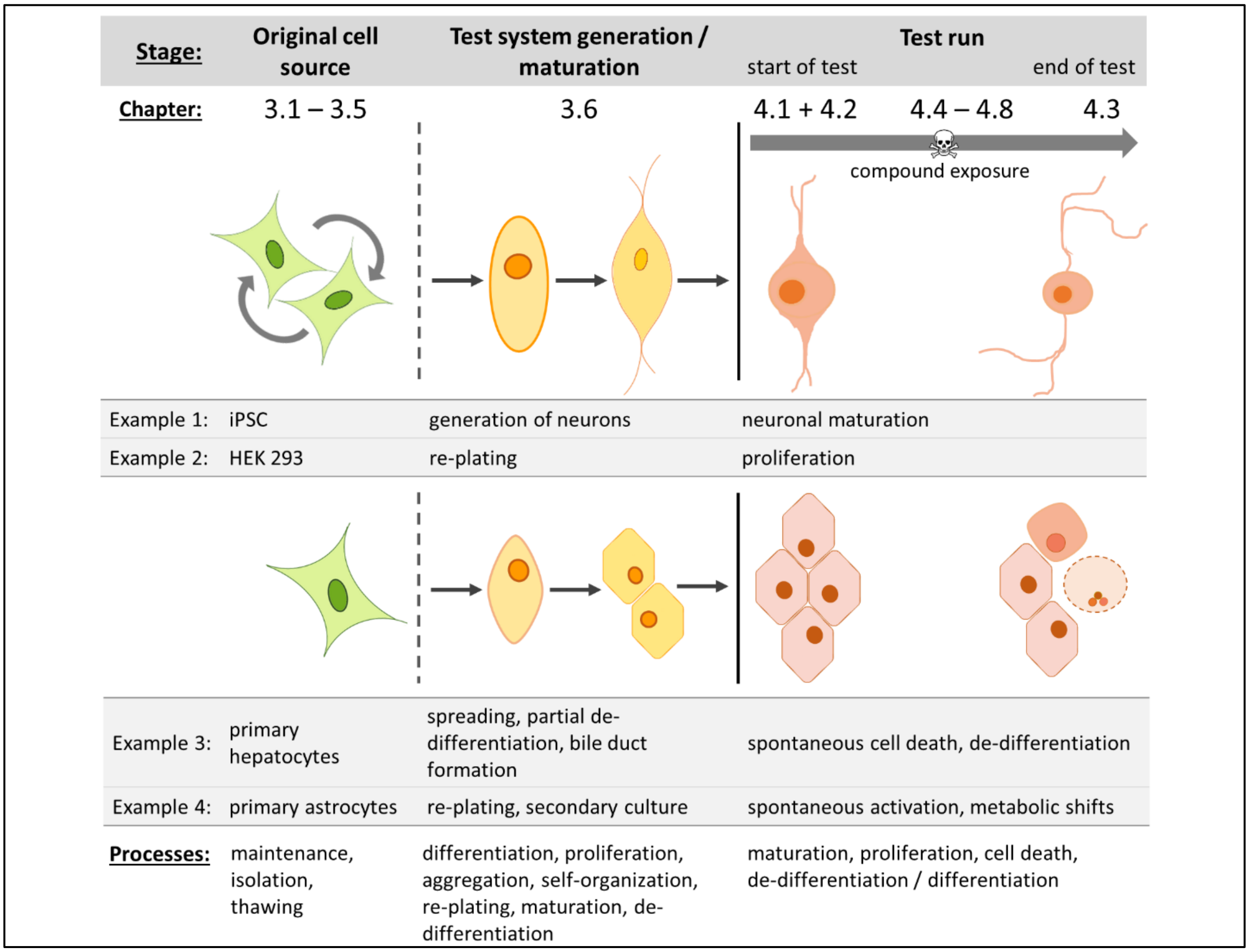
Supply number of supporting files.

Describe supporting files (e.g. metadata files, instrument settings, calculation template, raw data file, etc.).

3. Description of general features of the test system source

Note: *This section might be redundant with section 4. It is meant to describe the procedure of generating the cells, which are eventually used in the test method. This applies, e.g., for stem cells, which have to be differentiated towards the cell type with which the method is conducted (e.g., neurons or hepatocytes derived from iPSC). See scheme for illustration.*

If this is not applicable to your test system, go directly to section 4.



Stage:	**Original cell source**	**Test system generation / maturation**	**Test run** start of test … end of test
Chapter:	3.1 – 3.5	3.6	4.1 + 4.2 4.4 – 4.8 4.3
Example 1:	iPSC	generation of neurons	neuronal maturation
Example 2:	HEK 293	re-plating	proliferation
Example 3:	primary hepatocytes	spreading, partial de-differentiation, bile duct formation	spontaneous cell death, de-differentiation
Example 4:	primary astrocytes	re-plating, secondary culture	spontaneous activation, metabolic shifts
Processes:	maintenance, isolation, thawing	differentiation, proliferation, aggregation, self-organization, re-plating, maturation, de-differentiation	maturation, proliferation, cell death, de-differentiation / differentiation

Fig. 1: Overview of test system stages and where to find/deposit corresponding information in this document

Note: Refer to overview figure to connect question number and cell culture stages.

3.1 Supply of source cells

Describe briefly whether the cells are from a commercial supplier, continuously generated by cell culture, or obtained by isolation from human/animal tissue (or other).

Note: Not all stages may be relevant / applicable for some tests, e.g., commercial cells may be thawed and used immediately for a transporter or metabolism assay. In such cases, sections such as 3.6 may not need to be filled in.

3.2 Overview of cell source component(s)

Give a brief overview of your biological source system, i.e. the source or starting cells that you use.

Which cell type(s) are used or obtained (e.g. monoculture/co-culture, differentiation state, 2D/3D, etc.)?

If relevant, give human donor specifications (e.g. sex, age, pool of 10 donors, from healthy tissue, etc.).

3.3 Characterization and definition of source cells

List quantitative and semi-quantitative features that define your cell source/starting cell population. For test methods that are based on differentiation, describe your initial cells, e.g. iPSC, proliferating SH-SY5Y; the differentiated cells are described in section 4.

Define cell identity, e.g. by STR signature (where available), karyotype information, sex (where available and relevant), ATCC number, passage number, source (supplier), sub-line (where relevant), source of primary material, purity of the cells, etc.

Describe defining biological features you have measured or that are FIRMLY established (use simple listing, limit to max. 0.5 pages), e.g. the cells express specific marker genes, have specific surface antigens, lack certain markers, have or lack a relevant metabolic or transporting capacity, have a doubling time of x hours, etc.

Transgenic cell lines have particular requirements concerning the characterization of the genetic manipulation (type of transgene, type of vector, integration/deletion site(s), stability, etc.).

Organoids and microphysiological systems (MPS) may need some special/additional considerations as detailed in Pamies et al. (2018) and Marx et al. (2016), e.g. ratio of cell types used, percent of normal cells in tumor spheroids created from resected tissue; derivation of cells for re-aggregating brain cultures.

Note: Each test method can only use one specific cell source. Use of another source means that a key feature of the test method has been modified, and this change usually results in a new/different test method. (Examples: use of another transgene in a transgenic cell, or use of another cell line, or use of another mouse strain, or use of another primary cell isolation method are all likely to qualify for/require a new method description – including its characterization for performance parameters.

[“Source” is not limited to a single cell type, and it also includes composite sources, i.e., combination of ≥ 2 cell populations in a co-culture or microphysiological system. Notably, change of any part of such a composite source (e.g., one of the cell populations) would be considered a change of source)].

3.4 Acceptance criteria for source cell population

Describe the acceptance criteria (AC) for your initial cells (i.e. the quality criteria for your proliferating cell line, tissue for isolation, organism, etc.). Which specifications do you consider to describe the material, which quality control criteria have to be fulfilled (e.g. pathogen-free)? Which functional parameters (e.g. certain biological responses to reference substances) are important?

For iPSC maintenance: How do you control pluripotency? How stable are your cells over several passages? Which passage(s) are valid?

For primary cells: Show stability and identity of supply; demonstrate stability of function (e.g. xenobiotic metabolism).

Quantitative definitions for AC should be given based on this defining information. Exclusion criteria (features to be absent) are also important.

As in 3.3., special/additional requirements apply to genetically-modified cells and microphysiological systems.

3.5 Variability and troubleshooting of source cells

Name known causes of variability of the initial cells/source cells.

Indicate critical consumables or batch effects (e.g. relevance of the plate format and supplier, batch effects of fetal calf serum (FCS) or serum replacement, critical additives like type of trypsin, apo-transferrin vs. holo-transferrin, etc.).

Indicate critical handling steps and influencing factors (e.g. special care needed in pipetting, steps that need to be performed quickly, cell density, washing procedures, etc.).

As in 3.3., special/additional requirements apply to genetically-modified cells and microphysiological systems, e.g. dependence on matrix chemistry and geometry, dependence on microfluidics system, consideration of surface cells vs core cells, etc.

Give recommendations to increase/ensure reproducibility and performance.

3.6 Differentiation towards the final test system

Describe the principles of the selected differentiation protocol, including a scheme and graphical overview, indicating all phases, media, substrates, manipulation steps (medium change/re-plating, medium additives, etc.). Special/additional requirements apply to microphysiological systems and organoids: e.g. cell printing, self-aggregation/self-organisation process, interaction with the matrix, geometrical characterization (size/shape), etc.

Note: 'Differentiation' is meant here in a broad sense, involving all changes occurring to source cells until they are exposed to compounds. Examples for such processes include: proliferation associated with contact inhibition and change of average cell cycle state, formation of monolayers associated with formation of tight junctions, baso-lateral polarization, synaptic networking, formation of bile canaliculi, self-organization in 2D or 3D, change of activation state, etc. Thus, differentiation may be directed and intentional, but also non-intentional known changes need description here.

Examples to be described in detail: primary human hepatocytes (PHH) are plated on coated plastic, or are embedded in a collagen sandwich, or are aggregated to microtissues. Primary human monocytes are purified from blood by centrifugation and left to adhere on plastic; bronchial epithelial cells are cultured on air-liquid interface (ALI) until confluence is reached, etc.

3.7 Reference/link to maintenance culture protocol

Provide here the SOP of the general maintenance procedure as a database link. This should also include the following information:

How are the cells maintained outside the experiment (basic cell propagation)?

How pure is the cell population (average, e.g. 95% of iPSC cells Oct4-positive)?

What are the quality control measures and acceptance criteria for each cell batch?

Which number(s) passage(s) can be used in the test?

Is Good Cell Culture Practice (GCCP) and/or Good In Vitro *Method Practice (GIVIMP) followed?*

How long can same cell batches be used?

How are frozen stocks and cell banks prepared?

For primary cells: how are they obtained in general and what are they characterized for (and what are inclusion and exclusion criteria).

4. Definition of the test system as used in the method

Note: *This section refers to the stage of the test system, which is then used for the test method. See scheme for illustration. If you have cells that do not need prior differentiation, give their basic characteristics here. If your test system undergoes significant changes between the maintenance culture and the use for testing, please also fill in section 3.*

4.1 Principles of the culture protocol

Describe the test system as it is used in the test.

If the generation of the test system involves differentiation steps or complex technical manipulation (e.g. formation of microtissues), this is described in 3.6.

Give details on the general features/principles of the culture protocol (collagen embedding, 3D structuring, addition of mitotic inhibitors, addition of particular hormones/growth factors, etc.) of the cells that are used for the test.

What is the percentage of contaminating cells; in co-cultures what is the percentage of each subpopulation?

Are there subpopulations that are generally more sensitive to cytotoxicity than others, and could this influence viability measures? Is it known whether specific chemicals/chemical classes show differential cytotoxicity for the cell sub-populations used?

Note: The exact experimental procedure is covered by a separate question.

4.2 Acceptance criteria for assessing the test system at its start

What are the endpoint(s) that you use to control that your culture(s) is/are as expected at the start of toxicity testing (e.g. gene expression, staining, morphology, responses to reference chemicals, etc.)?

Describe the acceptance criteria for your test system, i.e. the quality criteria for your cells/tissues/organoids: Which endpoints do you consider to describe the cells or other source material, which parameters are important?

Note: To some extent, larvae/fish embryos may also be used in in vitro *methods, and a similar method description would apply.*

Describe the (analytical) methods that you use to evaluate your culture (PCR, ATP measurement) and to measure the acceptance criteria (AC).

Which values (e.g. degree of differentiation or cell density) need to be reached/should not be reached?

Historical controls: How does your test system perform with regard to the acceptance criteria, e.g. when differentiation is performed 10 times, what is the average and variation of the values for the acceptance criteria parameters)? Indicate actions if the AC are not met.

Examples: cell are > 90% viable, or > 98% of cells express marker x (e.g. AP-2), or > 80% of the cells attach, etc.

4.3 Acceptance criteria for the test system at the end of compound exposure

Note: Sometimes the test system does not change significantly between the beginning and end of compound exposure. In such cases, 4.3 does not need to be answered. In other cases, drastic changes occur (e.g. proliferation, further differentiation, spontaneous death, etc.). This needs characterization here.

Note: A common way to define 4.3 is to take data obtained for negative controls or solvent samples.

Describe the acceptance criteria for your test system, i.e. the quality criteria for your cells/tissues/organoids: Which endpoints do you consider to describe the cells or other source material, which parameters are important?

Which values (e.g. degree of differentiation or cell density) need to be reached/should not be reached?

Historical controls: How does your test system perform with regard to the acceptance criteria, e.g. when differentiation is performed 10 times, what is the average and variation of the values for the acceptance criteria parameters)? Indicate actions if the AC are not met.

Examples: Usual neurite length is 50 ±15 µm; experiments with average neurite length below 25 µm in the negative controls (NC) are discarded. Usual nestin induction is 200 ±40 fold, experiments with inductions below 80-fold for NC are discarded.

4.4 Variability of the test system and troubleshooting

Give known causes of variability for final test system state.

Indicate critical consumables or batch effects (e.g. plate format and supplier, batch effects of FCS or serum replacement, additives).

Indicate critical handling steps, and/or influencing factors identified (e.g. special care needed in pipetting, steps that need to be performed quickly, cell density).

Indicate positive and negative controls and their expected values, and accepted deviation within and between the test repeats.

Give recommendations to increase/ensure reproducibility and performance.

4.5 Metabolic capacity of the test system

What is known about endogenous metabolic capacity (CYP system (phase I); relevant conjugation reactions (phase II))?

What is known about other pathways relevant to xenobiotic metabolism?

What specific information is there on transporter activity?

Note: This paragraph is meant as a brief overview based on own experience and may differ from the general literature. In case of differences, information under 4.5 may, e.g., look like "normally these cells are reported to be devoid of CypIIA1 activity, but under our culture conditions, significant activity was observed.

4.6 Omics characterization of the test system

Are there transcriptomics data or other omics data available that describe the test system (characterization of cells without compounds)? Briefly list and describe such data.

Indicate the type of data available (e.g. RNASeq or proteomics data).

Refer to data file, data base or publication.

4.7 Features of the test system that reflect the *in vivo* tissue

Note: A differentiated cell or a cell line (such as HepG2) does not necessarily reflect all features of the corresponding in vivo *tissue / conditions.*

Give information on where the test system differs from the mimicked human tissue and which gaps of analogy need to be considered.

4.8 Commercial and intellectual property rights aspects of cells

Are there elements of the test system that are protected by patents or any other means?

Note: Here information can / should be added on the availability / accessibility of the test system (e.g. from a supplier or through a license agreement). This is also the place to mention potential limitations of use (imposed by the supplier or the intellectual property rights situation). Further guidance may be found at the OECD (http://www.oecd.org/chemicalsafety/testing/intellectual-property-in-oecd-testguidelines.htm) or in OECD GD 298 on availability of test system elements (Guiding Principles on Good Practices for the Availability/Distribution of Protected Elements in OECD Test Guidelines).

Note: Black boxes (missing/confidential information) are hardly acceptable for a fully valid test method description, and this situation is even more problematic for method validation (Linge and Hartung, 2007).

4.9 Reference/link to the culture protocol

Fill only if section 3 has not been answered.

Provide the SOP for the general maintenance procedure as a database link. This should also include the following information:

How are the cells maintained outside the experiment (basic cell propagation)?

How pure is the cell population (average, e.g. 95% of iPSC cells Oct4-positive)?

What are the quality control measures and acceptance criteria for each cell batch?

Which number(s) passage(s) can be used in the test?

Is Good Cell Culture Practice (GCCP) and/or Good In Vitro *Method Practice (GIVIMP) followed?*

How long can same cell batches be used?

How are freezing stocks and cell banks prepared?

For primary cells: How are they obtained in general and what are they characterized for (and what are inclusion and exclusion criteria).

5. Test method exposure scheme and endpoints

5.1 Exposure scheme for toxicity testing

Provide an exposure scheme (graphically, show timelines, addition of medium supplements and compounds, sampling, etc.), within the context of the overall cell culture scheme (e.g. freshly re-plated cells or confluent cells at start, certain coatings, etc.).

Include medium changes, cell re-plating, whether compounds are re-added in cases of medium change, critical medium supplements, etc.

Note: There can only be one exposure scheme, no alternative options! See 6.8 for indicating theoretical variations.

5.2 Endpoint(s) of the test method

Define the specific endpoint(s) of the test system that you use for toxicity testing (e.g. cytotoxicity, cell migration, etc.).

Indicate whether cytotoxicity is the primary endpoint.

What are secondary/further endpoints?

Also describe here potential reference/normalization endpoints (e.g. cytotoxicity, protein content, housekeeping gene expression) that are used for normalization of the primary endpoint.

5.3 Overview of analytical method(s) to assess test endpoint(s)

Define and describe the principle(s) of the analytical methods used. Provide here a general overview of the method's key steps (e.g. cells are fixed or not, homogenized sample or not, etc.), sufficient for reviewers/regulators to understand what was done, but not in all detail for direct repetition.

If you have two or more endpoints (e.g. viability and neurite outgrowth), do you measure both in the same well, under same conditions in parallel, or independently of each other?

For imaging endpoints: Explain in general how quantification algorithm or how semi-quantitative estimates are obtained and how many cells are imaged (roughly).

5.4 Technical details (of e.g. endpoint measurements)

Provide information on machine settings, analytical standards, data processing and normalization procedures.

For imaging endpoints: provide detailed algorithm.

This information should also be covered in an SOP, preferably in DB-ALM format (see link in 6.6).

Note: Details on data processing are given in a separate chapter below.

5.5 Endpoint-specific controls/mechanistic control compounds (MCC)

Note: MCCs are meant to control that your test method endpoint works, i.e., that it reacts to a biologically/mechanistically-relevant change as expected and that it does this every testing time and to a similar extent. They are also referred to as technical controls or as positive controls for the test method technical performance (Leist 2010).

MCC are chemicals/manipulations that show biologically plausible changes of the endpoint. List such controls (up to 10), indicate why you consider them as MCC, and describe expected data on such controls. Highlight the compounds to be used for testing day-to-day test performance, i.e. for setting acceptance criteria (AC).

If available, indicate MCC that each increase or decrease the activity of the relevant pathway. Do pathway inhibitions or activations correlate with the test method response?

Example 1: U0126 (ERK signaling pathway inhibitor). Neurite outgrowth in the CNS is controlled by ERK, inhibitors should therefore block this endpoint. U0126 blocks neurite growth at concentrations that block ERK activation

Example 2: Cytochalasin D (actin depolymerizer). Cell movement requires actin reorganization. Disturbance of actin structure should attenuate cell migration. Cytochalasin D inhibits cell migration at non-cytotoxic concentrations.

Example 3: BMP4 (endogenous protein, ligand of BMP receptor). Cell differentiation towards neuroectoderm is disturbed by BMP-SMAD signaling. Therefore, the test is based on SMAD inhibition by noggin (a protein scavenging/neutralizing BMP4). Addition of more BMP4 should outcompete noggin and lead to SMAD signaling, therefore preventing neuronal differentiation. BMP4 prevents the normal differentiation this test is based on.

5.6 Positive controls

Note: Positive controls (PC) are compounds that are known to affect the endpoint in man (or in a gold standard model). They are therefore also referred to as positive reference compounds. It is not necessary that the exact mechanism for the PC is known. However, for the use of a positive control (PC) in an in vitro *system, it is essential to show that the compound reaches the target cell* in vitro *similarly to the* in vivo *situation. (Example problem 1: The positive control (PC) reaches the target* in vivo *because of transport or because of accumulation not present* in vitro*; Example problem 2: The compound reaches and affects the target cell as a metabolite* in vivo*, and such metabolism is absent* in vitro*).*

What chemicals/manipulations are used as positive controls? Describe the expected data on such controls (signal and its uncertainty)?

How good are in vivo *reference data on the positive controls? Are* in vivo *relevant threshold concentrations known?*

Note: MCC and PC may overlap. MCC are often more suitable than PC in order to define acceptance criteria for the test method; PC are classically used to build the prediction model.

Note: PC (and other chemicals used as anchor for the test method performance) need specification (not just listing of names), e.g., their CAS number, purity, handling and storage (relevant also for 5.5 and 5.7).

5.7 Negative and unspecific controls

Note: Negative controls (NC) for the prediction model (PM) are compounds or substances known to NOT affect the endpoint in vivo *(e.g. folic acid for developmental neurotoxicity (DNT) endpoints, as folic acid is recommended during pregnancy). They may also be termed negative reference compounds.*

Negative controls for the technical performance of the test method (NC_{TM}) are used to set acceptance criteria. They are compounds that do not change the normal (undisturbed) readout significantly. This may include solvent controls. Often NC can be used as NC_{TM}.

An unspecific control (UC) is a compound that has activity, e.g. is cytotoxic, but does not affect the functional (main) readout of an assay (Leist 2010). UC are absolutely essential to define baseline variation of functional tests, and thus to build their prediction model (PM).

What chemicals/manipulations are used as negative controls? Describe the expected data on such controls (signal and its uncertainty)? (Such data define the background noise of the test method)

What is the rationale for the concentration setting of negative controls?

Do you use unspecific controls? If yes, indicate the compounds and the respective rationale for their use and the concentration selection.

Note: Some ways to define negatives include: (i) compound only acting when metabolized, (ii) acting on another organ, (iii) known to be safe for man, (iv) being selective for another assay, (v) pairs/matches of a specific positive control (e.g. inactive metabolite). However, this all needs background knowledge. See also 5.6 for toxicokinetics problems (Example problems: A compound may not affect a target cell in vivo *because it does not reach the cell behind a barrier or the compound has a too short half-life (high excretion or metabolism).*

5.8 Features relevant for cytotoxicity testing

Does the test system have a particular apoptosis sensitivity or resistance?

Is cytotoxicity hard to capture for minor cellular subpopulations?

In multicellular systems, which cell population is the most sensitive? Are specific markers known for each cell population?

Are there issues with distinguishing slowed proliferation from cell death?

For repeated/prolonged dosing: Is early death and compensatory growth considered?

For very short-term endpoints (e.g. electrophysiology measured 30 min after toxicant exposure): Is a delayed measure of cytotoxicity provided?

Note: This paragraph is meant as a brief overview based on own experience; it may differ from the general literature. In case of differences, these should be specified.

5.9 Acceptance criteria for the test method

Note: Acceptance criteria decide when a test run is discarded.

It is recommended to set acceptance criteria, e.g., value ±SD for such controls. The values should be recorded over time in the "historical control data". Historical control data can be used as additional control to ensure that the test system is working adequately (not deviating in its performance over time).

Which rule do you apply to test whether a test run is within the normal performance frame?

How do you document this decision?

Indicate actions if the AC are not met.

Note: AC are usually defined on the base of PC or MCC or NC_{TM} results run in parallel.

MCC or NC_{TM} have to meet, for example, certain threshold values or be in a historic range to accept the given test run. As an example, such reference compounds were discussed for developmental neurotoxicity (DNT) by Aschner et al. (2017).

Note: Reference materials (MCC, or NC_{TM}) also play a central role to show the proficiency of a laboratory to perform a test, or they can be used to document the adequacy of a certain test implementation/variant compared to a (validated) reference method (Hartung, 2007).

5.10 Throughput estimate

Indicate "real data points per month" (not per week/per quarter, etc.): count three working weeks per month. Each concentration is a data point. Necessary controls that are required for calibration and for acceptability criteria are NOT counted as data points. All technical replicates of one condition are counted as one single data point (see notes for explanation)

Note: Reference materials (MCC or NC_{TM}) also play a central role in showing the proficiency of a laboratory to perform a test, or they can be used to document the adequacy of a certain test implementation/variant compared to a (validated) reference method (Hartung, 2007).

Indicate possibility/extent of repeated measures (over time) from same dish.

Explain your estimate.

Note: The throughput estimate refers to the test conditions available to the developer or the lab where the test is applied. It refers to the normal required number of replicates needed to obtain meaningful data (as specified by the prediction model).

Note: Technical replicates are several measurements of the same sample, e.g., five solvent samples on the same multi-well plate or in the same test run, using cells from the same cell batch. Biological replicates are when the test is repeated on another day or using different cells (different passage or differentiation), see also Leist et al. (2010).

6. Handling details of the test method

6.1 Preparation/addition of test compounds

Give an overview of the range of volumes, particular lab ware/instruments for dispensing, temperature/lighting considerations, particular media/buffers for dilution, decision rules for the solvent, tests of solubility as stocks and in culture medium, etc.

How are compound stocks prepared (fold concentration, verification, storage, etc.)?

How are dilutions prepared? What solvent is used? Is filtering used to obtain sterility?

How does the final addition to the test system take place?

Give details of addition of test compounds to test systems (e.g. in which compartment of compartmentalized cultures, in which volume, before after or during medium change, etc.).

Note: Strike a reasonable balance between information overview for reviewers (focus here) and technical detail (to be detailed in an SOP). This applies to several chapters detailed in an SOP (e.g., controls and acceptance criteria).

6.2 Day-to-day documentation of test execution

How are day-to-day procedures documented (type of 'lab book' organisation, templates)?

Define lab-specific procedures used for each practical experiment on how to calculate test compound concentrations (and to document this).

How are plate maps defined and reported?

Detailed information should also be covered in an SOP, preferably in DB-ALM format (see link in 6.6).

Note: Lab book entries are metadata that need to be linkable to test results on demand!

6.3 Practical phase of test compound exposure

How is the time plan of pipetting established, followed, and documented?

How is adherence to plate maps during pipetting documented?

What are the routine procedures to document intermediate steps with potential errors, mistakes and uncertainties?

How are errors documented (e.g. pipetting twice in one well)?

How are the plate wells used sequentially – following which pattern?

Detailed information should also be included in an SOP, preferably in DB-ALM format (see link in 6.6).

6.4 Concentration settings

How is the concentration range of test compounds defined (e.g. only single concentrations, always 1:10 serial dilutions or variable dilution factors, ten different concentrations, etc.)? Is there a rule for defining starting dilutions?

For functional endpoints that may not provide full concentration-response, how is the test concentration defined? E.g. EC_{10} of viability data are usually tested for gene expression endpoints.

Note: For concentration-dependent data, no-effect concentrations must be included (full range curve). Data need sufficiently dense spacing around benchmark concentration; preferably provide statistical significance for key data points.

Detailed information should also be included in an SOP, preferably in DB-ALM format (see link in 6.6).

6.5 Uncertainties and troubleshooting

What types of compounds are problematic, e.g. interference with analytical endpoint, low solubility, precipitation of medium components, etc.?

What experimental variables that are hard to control (e.g. because they are fluorescent)?

What are critical handling steps during the execution of the assay?

Robustness issues, e.g. known variations of test performance due to operator training, season, use of certain consumable or unknown causes, etc.

Describe known pitfalls (or potential operator mistakes).

6.6 Detailed protocol (SOP)

Ideally the SOP follows the DB-ALM or a comparable format:

https://ecvam-dbalm.jrc.ec.europa.eu/home/contribute

Refer to additional file(s) (containing information covered in sections 3 and 4), containing all details and explanations.

Has the SOP been deposited in an accessible data base?

Has the SOP been reviewed externally and if yes, how?

6.7 Special instrumentation

Does the method require specialized instrumentation that is not found in standard laboratories?

Is there a need for custom-made instrumentation or material?

Is there a need for equipment that is not commercially available (anymore)?

6.8 Possible variations

Note: As mentioned above, variations of the main elements of an assay usually require that a new assay is defined and characterized. Such new assays may be highly related to the one described here. Sometimes method descriptions for the related assays may not be available, but data from them may be found in the literature or be used for comparison with data from the test method described here. In such cases, this chapter is the place to list such related assays and to describe the element that differs.

Describe possible variations, modifications and extensions of the test method:

a) other endpoints,

b) other analytical methods for same endpoint,

c) other exposure schemes (e.g. repeated exposure, prolonged exposure, etc.),

d) experimental variations (e.g. use of a specific medium, presence of an inhibitor or substrate that affects test outcome, etc.)

Note: This should not be everything that COULD be done, but only points that have been done successfully, or that have a high likelihood of being done in a defined project context and after some appropriate evaluation.

6.9 Cross-reference to related test methods

Indicate the names (and database names) of related tests and give a short description (including a brief comment on differences to the present method).

If the test method has been used for high throughput transcriptomics or deep sequencing as alternative endpoint, this should be indicated.

7. Data management

7.1 Raw data format

What is the data format?

Raw data: give general explanation. Upload an exemplary file of raw data (e.g. Excel file as exported out of plate reader).

Provide an example of processed data at a level suitable for general display and comparison of conditions and across experiments and methods.

Note: It is recommended that data formats suitable for most/all methods are pre-defined in collaborative projects, such as EU-ToxRisk or ToxCast.

If the file format is not proprietary or binary, include a template. This will help other users to provide their data in a similar way to the general data infrastructure.

Example as used in EU-ToxRisk: Excel sheet with columns specifying line number, assay name, date of experiment, identifier for reference to partner lab book, compound, concentration (in: -log[M]), line number of corresponding control, number of replicates, endpoints, data of endpoint(s), etc.

7.2 Outliers

How are outliers defined and handled?

How are they documented?

Provide the general frequency of outliers.

Note: If only processed data are reported, outlier information gets lost.

7.3 Raw data processing to summary data

How are raw data processed to obtain summary data (e.g. EC_{50}, BMC15, ratios, PoD, etc.) in your lab?

Describe all processing steps from background correction (e.g. measurement of medium control) to normalization steps (e.g. if you relate treated samples to untreated controls).

7.4 Curve fitting

How are data normally handled to obtain the overall test result (e.g. concentration response fitting using model X, determination of EC_{50} by method Y, use of EC_{50} as final data)?

How do you model your concentration response curve (e.g. LL.4 parameter fit) and which software do you use (e.g. GraphPad Prism, R, etc.)?

Do you usually calculate an uncertainty measure of your summary data (e.g. a 95% confidence interval for the BMC or a BMCL), and with which software?

Can you give uncertainty for non-cytotoxicity or no-effect?

How do you handle non-monotonic curve shapes or other curve features that are hard to describe with the usual mathematical fit model?

7.5 Internal data storage

How and how long are raw and other related data stored?

What backup procedures are used (how frequently)?

How are data versions identified?

7.6 Metadata

Note: Metadata are, for example, laser power, microscope objective, binning of camera, slit/filter of optical units, temperature cycle of PCR, all data that refer to instrument settings during data recording, suppliers of chemicals, software versions for data processing, types of dishes, media and consumables used, etc.

How are metadata documented and stored (lab book, Excel files, left in machine, etc.)?

How are they linked to raw data?

What metadata are stored/should be stored?

7.7 Metadata file format

Give example of the metadata file (if available).

Note: Also consider to include fields which are variable, but cannot be completely specified in the protocol and could be changed without changing the readout (e.g. suppliers of chemicals) in this template.

If metadata or data format (see 7.1) are pre-defined in the project, state here "as pre-defined in project xxx" (e.g. EU-ToxRisk).

8. Prediction model and toxicological application

8.1 Scientific principle, test purpose and relevance

What is the scientific rationale to link test method data to a relevant in vivo *adverse outcome?*

Note: The following questions further specify this point.

Which toxicological target (organ, tissue, physiological/biochemical function, etc.) is modelled?

Which biological process(es) (e.g. neurite outgrowth, differentiation) are modelled/reflected by your test method?

Which toxicological events (e.g. oxidative stress, cell death) are modelled/reflected by your test method?

To which (human) adverse outcome(s) is your test method related?

Note: The method description may apply equally to human toxicology (main objective here) and to animal and ecotoxicological test methods.

Which hazard(s) do(es) your test method (potentially) predict?

8.2 Prediction model

Provide the statistics of your benchmark response (threshold and variance):

(i) For dichotomized data, provide your prediction model. When do you consider the result as toxic or not toxic?

(ii) For pseudo-dichotomized outcomes (two classes with borderline class in between): define borderline range.

(iii) For multi-class or continuous outcomes: provide definitions and rationale.

Note: Option (iii) is uncommon and hard to use in practice: requires very good rationale!

What is the rationale for your threshold? This can be on a mathematical (e.g. 3-fold standard deviation) or a biological basis (e.g. below 80% viability).

Is there a toxicological rationale for the threshold settings and definitions of your prediction model?

Note: There may be different prediction models for different regulatory applications. For each application/prediction model, a rationale (weight-of-evidence analysis) should be given.

What are the limitations of your prediction model?

What is a 'hit' if the test is used in screening mode (= hit definition, if different from above)?

Note: Hit definition in a screen can be different from the benchmark response in normal testing, e.g., in a screen you can be less strict because you do not want to miss anything, whereas in hit follow up testing you may use the benchmark response threshold.

8.3 Prediction model setup

How was the prediction model set up (using which test set of chemicals to train the model; using probing with what kind of classifiers/statistical approaches)?

Has the prediction model been tested (what was the test set of chemicals)? List chemicals or give n, if n > 50.

Is the process documented (publication)?

Does the prediction model (PM) apply to changes to both sides of controls (up/down)? If the PM is one-sided (e.g. toxicants leading to a decrease vs. control), how are data in the opposite direction handled and interpreted? If the PM is two-sided, do different rules, characteristics and interpretations apply to the two sides (e.g. is a decrease in viability or an increase in viability both interpreted as an effect/toxicity; are thresholds and performance characteristics to both sides the same?).

Note: For final values on accuracy, sensitivity and specificity, refer to 8.4.

8.4 Test performance

Indicate here basic performance parameters or, if possible, preliminary estimates (label as such): Baseline variation (noise) within assays AND between assays.

What is the signal/noise ratio (signal = standard positive control)?

Is the z-factor determined?

Give the specificity of the test method. How is it determined?

Give the sensitivity of the test method. How is it determined?

Give measures of the uncertainty of your test method. How are they determined?

What is the detection limit (required change of endpoint to become measurable)?

If available, give limit of detection (LOD) and limit of quantification (LOQ).

What are inter-operator variations?

Are there data of 'historical controls' over a longer time period?

Note: Assay parameters can only apply to one assay version. A standard version must be defined and referred to in all answers.

8.5 *In vitro – in vivo* extrapolation (IVIVE)

Describe parameters important for the determination of free compound concentrations in the medium.

Indicate the lipid and protein content of the medium and the cells.

Indicate the volume of the cells.

Indicate volume (medium volume) and surface area of culture dish.

Is there information/literature on IVIVE strategies/data in the test?

Has the test been used earlier for IVIVE?

Are there special considerations that are relevant for IVIVE (e.g. potential for compound accumulation due to frequent medium changes and compound re-addition, glycoprotein (MDR1) expression, capacity for xenobiotic metabolism of test system)?

8.6 Applicability of test method

Note: This refers to the biological and chemical applicability.

Which compounds is the test likely to pick up correctly, where is it likely to fail?

How does the test method react to mixtures and UVCBs?

Are there areas (according to industry sector, compound chemistry, physical-chemical properties) that need to be excluded from testing, or that are particularly suitable?

Which compound class cannot be detected (e.g. neurotransmitters for which the receptors are not expressed, endocrine disruptors in absence of respective pathway)?

Are any compounds known to interfere with the test system (e.g. fluorescent or colored chemicals)?

8.7 Incorporation in test battery

Does the test fit into a test battery? If yes, into which test battery and are there any restrictions?

Indicate potential strengths and weaknesses of the system in a test battery (e.g. method is a good confirmation assay, good for creating alerts, mechanistic follow-up, screening, etc.).

Compare performance to similar tests.

Which gaps in a known or potential battery does the test method fill?

Should the test preferentially be used in the first tier or later tiers, are complementary assays required or is it a stand-alone method?

Note: Such information may not be available at early phases of test implementation and use. If it is available, it is a valuable part of the test description, as it specifies one of the test purposes and it links the test to related tests, with overall battery outcome as secondary purpose.

9. Publication/validation status

9.1 Availability of key publications

Refer to published literature on the test AND indicate in detail deviations from published descriptions (e.g. plastic plate supplier, cell number, endpoint measurement, timing, etc.).

Note: Provide links (e.g., to PubMed, doi, etc.).

Note: Reference for SOP under section 6.6.

Provide the most relevant publications that describe/give a comprehensive overview of (a) your test system and/or (b) your test method. Describe what aspects are covered therein.

Give a prioritized (according to importance) list of further publications on the test method or its application.

Give short comments on which type(s) of information can be obtained from these publications (e.g. contains test chemical lists, contains more positive/negative controls, contains validation against other tests, contains incorporation in test battery, demonstrates use by other lab, etc.).

9.2 (Potential) linkage to AOPs

Indicate whether the test method has been or could be linked to an AOP (or several AOPs) and in which form (e.g. test of KE activation).

Can the test method cover an AOP MIE/KE?

Reference relevant AOP and if in AOP-wiki, refer to status.

Note: See remarks in 8.7.

9.3 Steps towards mechanistic validation

Indicate/summarize information on mechanistic validation, e.g. by omics approaches or by use of endpoint specific controls (MCC; section 5.5).

Has it been explored in how far the system reflects human biology, signaling, tissue organization relevant to the form of toxicity to be assessed (e.g. nigrostriatal neurons should contain dopamine, liver tests relevant to cholestasis may need to contain bile canalicular structures, etc.)?

Are there interventions (knock-out, knockdown, chemical inhibitors, specific pathway triggering) that support the use of the test for certain toxicological questions and that corroborate expectations to the test system?

Is there a form of mechanistic validation?

Do toxicant-altered genes (or other biomarkers) correspond to changes in mimicked human tissue (after poisoning or in relevant pathologies)?

Example 1: If a test measures neurite growth, then biological signals known to control neurite growth and growth cone collapse should be present in the system and their modulation should affect the test endpoint.

Example 2: If a test measures DNA damage response, then DNA damage sensors should be expressed and functioning, and knockdown of DNA damage sensors should affect the test endpoint.

9.4 Pre-validation or validation

Indicate/summarize activities for test qualification, pre-validation or validation.

Indicate e.g. ring trials, full (pre-)validations.

Give an overview of compounds or libraries that have been tested.

Note: To give evaluators/regulators an overview of the test readiness, this may be the place to add a summary of test readiness scoring. This would give an overall overview of the uncertainties associated with test readiness (different from the single uncertainty measures for test system, analytical endpoint, the prediction model, etc.). Ideally, here the measures to minimize the impact of overall test method uncertainty may be described.

9.5 Linkage to (e.g. OECD) guidelines/regulatory use

Indicate whether the test method is linked to an OECD Test Guideline (how, and which) or other regulatory guidance (e.g. EMA).

10. Test method transferability

10.1 Operator training

What experience is required?

How are new operators trained in your laboratory?

How much training/experience is required for smooth assay performance?

10.2 Transfer

Has the test system been transferred to other labs?

Has the test method been used by various operators (over a long time period)?

Has the test method been transferred to other labs?

Is there data on inter-laboratory variability?

What are procedures and how was the performance (experience) of the transfer?

11. Safety, ethics and specific requirements

11.1 Specific hazards; issues of waste disposal

Are there special legal requirements for running the test in your lab; are there special hazards associated with the test that may affect operators, bystanders, others (e.g. through waste).

11.2 Safety data sheet (SDS)

Are the SDSs for all hazardous reagents used in the test method available?

Are the SDSs for all hazardous test compounds stored?

Describe where and how the SDSs are stored internally. How is safe handling ensured?

Is the exposure scenario for the hazardous reagents used in the test method available?

11.3 Specific facilities/licenses

Are special permits (e.g. genetic work, stem cells, radioactivity, etc.) required?

Are special facilities required?

Is special ethical approval necessary (indicate approval document).

11.4 Commercial aspects/intellectual property of material/procedures

List elements of the test method (e.g. consumables, chemicals, analytical methods, equipment) that are protected by patents or any other means. Indicate the type of protection and where the element (or license for it) may be obtained.

Note: Here information can/should be added on the availability/accessibility of the test elements (e.g. from a supplier or through a license agreement. This is also the place to mention potential limitations of use (imposed by the supplier or the intellectual property rights situation). Further guidance may be found at the OECD (http://www.oecd.org/chemicalsafety/testing/intellectual-property-in-oecd-testguidelines.htm). Compare also 4.8 (on commercial aspects of the cell system).

Note: Sometimes such information is hard to obtain. Unless full OECD validation and acceptance is sought, information may be given only on issues and elements that are well-known (or that may be a problem upon test method transfer).

S1.3 Abbreviations

AC	Acceptance criteria
ALI	air-liquid interface
AO	adverse outcome
AOP	adverse outcome pathway
ATCC	American Type Culture Collection
ATP	adenosine triphosphate
BMC	benchmark concentration
BMCL	benchmark concentration lower limit
BMP4	bone morphogenetic protein 4
CYP	Cytochrome P450

DB	database
DB-ALM	DataBase service on Alternative Methods to animal experimentation
DNT	developmental neurotoxicity
doi	digital object identifier
ECVAM	European Center for the Validation of Alternative Methods
EMA	European Medicines Agency
ESC	embryonic stem cell
FCS	fetal calf serum
GCCP	Good Cell Culture Practices
GIVIMP	Good *In Vitro* Method Practices
iPSC	induced pluripotent stem cell
IVIVE	*in vitro* – *in vivo* extrapolation
KE	key event
LL.4	four-parameter log-logistic function
LOD	limit of detection
LOQ	limit of quantification
M	molar
MCC	mechanistic control compounds
MIE	molecular initiating event
MPS	microphysiological system
MSDS	material safety data sheet
NC	negative control
Oct4	transcription factor, embryonic stem cell marker
OECD	organization for economic cooperation and development
PC	positive control
PCR	polymerase chain reaction
Pgp	permeability glycoprotein
PHH	primary human hepatocytes
PI	principal investigator
PoD	point of departure
SOP	standard operating procedure
UVCB	Unknown or Variable composition, Complex reaction products or Biological materials

S1.4 Comparison of the OECD Guidance Document 211 vs. ToxTemp

Chapter in GD211	Chapter name in GD 211	Specification on information to be provided	Chapter in ToxTemp	Chapter name in ToxTemp	Comment
1.	**General information**				
1.1	Assay Name (title)	Short and descriptive title	1.1	Descriptive full-text title	
1.2	Summary	Summary of assay features	1.2	Abstract	
1.3	Date of MD	Date of first version (D/M/Y)	2.2	Version number and date of deposition	
1.4	MD author(s) and contact details	Names	2.6	Name and email of contact person	
		Contact details	2.7	Name of further persons involved	
1.5	Date of MD update(s) and contacts	Date (D/M/Y) of update	2.2	Version number and date of deposition	
		Update can be for addition of new information or correction			
		Summary what has been updated	2.3	Summary of introduced changes in comparison to previous version(s)	
1.6	Assay developer(s)/Laboratory and contact details	Name of developer/lab/author	2.5	Name and acronym of the test depositor	
		Contact details			
1.7	Date of assay development and/or publication	Year of initial assay release/publication	2.2	Version number and date of deposition	
		Existence of potential public SOP			
1.8	Reference(s) to main scientific papers	List of bibliographic references to papers that explain assay development	9.1	Availability of key publications	
		References to e.g. validation datasets or prediction model should go to 6.0			

1.9	Availability of information about the assay in relation to proprietary elements	Is assay proprietary or non-proprietary?	4.8	Commercial and intellectual property rights aspects of cells	
			11.4	Commercial aspects / intellectual property of material / procedures	
		Is assay method transferable?	10.2	Transfer	Disclosure of information is not considered.
		Which information cannot be disclosed?			
1.10	Information about the throughput of the assay	Estimation of resource intensity	5.10	Throughput estimate	
		Give approx. no of compounds/concentrations per run			
1.11	Status of method development and uses	i) Development status: still under development, when finished?			Not stated if complete or not
		ii) Known uses: used in different laboratories?	10.2	Transfer	
		iii) Evaluation study: summarize main conclusions	1.2	Abstract	
			9.4	Pre-validation or validation	
		iv) Validation study: part of formal validation study? Summarize conclusion and outcomes	9.4	Pre-validation or validation	
		v) Regulatory use: details about regulatory application	9.5	Linkage to (e.g. OECD) guidelines / regulatory use	
			1.2	Abstract	
		Toxicological hazard endpoint being addressed	1.2	Abstract	
			8.1	Scientific principle, test purpose and relevance	
1.12	Abbreviation and Definitions				Included at the end
2.	**Test Method Definition**				
2.1	Purpose of the test method	What is the claimed purpose and rationale for intended use of method, e.g. alternative to an existing method, screening, provision of novel information in regulatory decision-making, mechanistic information, adjunct test, replacement, etc.	8.7	Incorporation in test battery	
			1.2	Abstract	
			8.1	Scientific principle, test purpose and relevance	
		What is relation between assay-response and *in vivo*/biology/physiology?	4.7	Features of the test system that reflect the *in vivo* tissue	
			8.1	Scientific principle, test purpose and relevance	
			9.3	Steps towards mechanistic validation	
		Link to KE or MIE?	9.2	(Potential) linkage to AOPs	
			9.3	Steps towards mechanistic validation	
		Which AO might be modelled?	9.2	(Potential) linkage to AOPs	
		If no AOP present, give link between mechanism the assay measures and resulting hazard endpoint	9.3	Steps towards mechanistic validation	
2.2	Scientific principle of the method	Provide scientific rationale	8.1	Scientific principle, test purpose and relevance	
		Description of scientific principle	8.1	Scientific principle, test purpose and relevance	
		Biological/physiological basis and relevance	4.7	Features of the test system that reflect the *in vivo* tissue	
			9.3	Steps towards mechanistic validation	
		Mechanistic basis	5.5	Endpoint-specific controls / mechanistic control compounds (MCC)	
		Is anchor point an AOP?	9.2	(Potential) linkage to AOPs	

2.3	Tissue, cells or extracts utilised in the assay and the species source	What is the experimental system for activity or response measured?	3.2	Overview of cell source component(s)	
		Is material commercially available?	3.1	Supply of source cells	
		Is material developed in lab?	3.1	Supply of source cells	
		Source/manufacturer of biological material	3.1	Supply of source cells	
		Can material be cryopreserved or freshly prepared?	3.2	Overview of cell source component(s)	
			3.1	Supply of source cells	
2.4	Metabolic competence of the test system	Is test system metabolically competent?	4.5	Metabolic capacity of the test system	
		Addition of enzymatic fraction?			
2.5	Description of the experimental system exposure regime	Summary description of exposure regime (dosage, exposure time, readout frequency)	5.1	Exposure scheme for toxicity testing	
		Number of doses/concentrations, testing range	6.4	Concentration settings	
		Number of replicates	5.10	Throughput estimate	
		Use of controls and vehicles	5.6	Positive controls	
			5.7	Negative and unspecific controls	
		Specialized equipment needed	6.7	Special instrumentation	
			11.3	Specific facilities / licenses	
		Potential solubility issues with the test system, and solutions proposed to address the issue	8.6	Applicability of test method	Variability and troubleshooting asked several times.
2.6	Response and Response Measurement	Response here makes reference to any biological effect, process or activity that can be measured	5.2	Endpoint(s) of the test method	
			1.2	Abstract	
			8.1	Scientific principle, test purpose and relevance	
			9.3	Steps towards mechanistic validation	
		Describe the response and its measurement	5.2	Endpoint(s) of the test method	
			5.3	Overview on analytical method(s) to assess test endpoint(s)	
		Specify precise response as applicable, e.g. IC50	7.3	Raw data processing to summary data	
		Description how it is calculated	7.4	Curve fitting	
2.7	Quality / Acceptance criteria	Information on the availability of acceptance criteria and quality assurance	3.4	Acceptance criteria for source cell population	Acceptance criteria assessed at different stages of test system and method, furthermore assessing variability and troubleshooting.
			4.2	Acceptance criteria for assessing test system at its start	
			4.3	Acceptance criteria for the test system at the end of compound exposure	
		Experimental data (storage/archiving), give unit of raw data	7.1	Raw data format	
			7.5	Internal data storage	
		Experimental system(s) used	3.2	Overview of cell source component(s)	Experimental test system

					extensively documented at different stages.
		Equipment used, calibration program	5.4	Technical details (of e.g. endpoint measurements)	
			6.7	Special instrumentation	
		Availability of internal standards (e.g. positive and negative controls)	5.5	Endpoint-specific controls / mechanistic control compounds (MCC)	
			5.6	Positive controls	
			5.7	Negative and unspecific controls	
		Standards followed such as good cell culture practice	9.5	Linkage to (e.g. OECD) guidelines / regulatory use	
		Criteria to accept or reject experimental data	4.3	Acceptance criteria for the test system at the end of compound exposure	Also look at acceptance criteria for test system at different stages.
		Limit of detection and limit of quantification, detection range	8.4	Test performance	
2.8	Known technical limitations and strengths	Specification of any known technical limitations or strengths of running assay	5.8	Features relevant for cytotoxicity testing	
			8.6	Applicability of test method	
			8.7	Incorporation in test battery	
		The assay may not be technically applicable to certain types or class of chemicals.	8.6	Applicability of test method	
2.9	Other related assays that characterise the same event as in 2.1	Related assays if known that may characterise the same key event as described in 2.1	6.9	Cross-reference to related test methods	
3.	**Data interpretation and prediction model**				
		If applicable, give brief summary and references for the prediction model	8.2	Prediction model	
		Prediction model might arise from assay, or battery	8.3	Prediction model setup	
		Consider the intended purpose of the prediction model	8.2	Prediction model	
		Specify if this refers to key events as defined in AOPs	9.2	(Potential) linkage to AOPs	
			9.3	Steps towards mechanistic validation	
3.1	Assay response(s) captured in the prediction model	Identify the response(s) from the given assay(s) that form(s) the basis of the prediction model	5.2	Endpoint(s) of the test method	
			8.2	Prediction model	
3.2	Data analysis	Comment on the response value in terms of a boundary or range to provide a context for interpretation.	8.4	Test performance	
3.3	Explicit prediction model	Description of prediction model	8.2	Prediction model	Also test performance included.
3.4	Software name and version for algorithm/prediction model generation	Software used to derive the prediction model or to undertake the statistical processing.	7.4	Curve fitting	
4.	**Test Method Performance**				
4.1	Robustness of the method	Reliability of the experimental results	8.4	Test performance	
		Within-laboratory repeatability and reproducibility	8.4	Test performance	
			10.2	Transfer	
		Between laboratory transferability and reproducibility	10.2	Transfer	
4.2	Reference chemicals/chemical libraries, rationale for	Are results for the reference chemicals (i.e. the “training set” chemicals used in the	8.3	Prediction model setup	

	their selection and other available information	development and evaluation of the assay and its associated prediction model) are free and publicly available in some form			
		If available, what is rationale for their selection	5.5	Endpoint-specific controls / mechanistic control compounds (MCC)	
			5.6	Positive controls	
			5.7	Negative and unspecific controls	
		Give all information available on compound, e.g. chemical names, CAS, SMILES, structure, InCHI code, etc.			Not requested, as tested compounds do not belong to test method. Positive, negative and unspecific controls should be given.
		For mixtures, report the composition	8.6	Applicability of test method	
4.3	Performance measures/predictive capacity (if known)	Give goodness-of-fit statistics or goodness-of-fit testing (e.g. r2, r2 adjusted, standard error, sensitivity, specificity, false negative and false positive rates, predictive values)			Statistical goodness-of fit is not requested, but uncertainty of summary data is asked.
		Rationale for application of certain function	7.4	Curve fitting	
		Specification of the fit	7.4	Curve fitting	
		Explanation of the curve fitting process	7.4	Curve fitting	
		Limitations related to the data analysis	7.4	Curve fitting	
		Was cross-validation carried out and statistics used?	8.3	Prediction model setup	
			9.4	Pre-validation or validation	
4.4	Scope and limitations of the assay, if known	Types of substances for which the assay is appropriate	8.6	Applicability of test method	
		Applicability domain based on molecular descriptors	8.6	Applicability of test method	Molecular descriptors are not specified.
		Physical-chemical limitations of compounds	8.6	Applicability of test method	
		Test amenable to variety of chemicals such as mixtures, UVCBs, multi-constituent substances, organometallics, inorganic substances, discrete organic substances and various chemical classes or organic substances?	8.6	Applicability of test method	
		What are inclusion and/or exclusion rules for compounds, e.g. volatility	8.6	Applicability of test method	
		Indications from the false positives/false negatives identified that the assay has specific limitations?	8.2	Prediction model	Limitation of prediction model
5.	**Potential Regulatory applications**				
		Build a contextual weight of evidence analysis on the use of the prediction model for different regulatory purposes, indicate all its potential applications	8.1	Scientific principle, test purpose and relevance	
5.1	Context of use	Possible conditions of use?	8.6	Applicability of test method	

		Give scientific confidence for different end use scenarios	8.1	Scientific principle, test purpose and relevance	
			8.4	Test performance	
		Scientific confidence for the use of a given prediction model and the rationale for this	8.2	Prediction model	
			8.4	Test performance	
		Possible end use scenarios	8.7	Incorporation in test battery	
		Support category formation and read-across			Not covered
		Priority setting	8.7	Incorporation in test battery	
		Screening level assessment	8.7	Incorporation in test battery	
		Integrated approaches to testing and assessment (IATA)	8.7	Incorporation in test battery	
6.	**Bibliography**				
		Useful references not associated with assay or prediction model development	9.1	Availability of key publications	
7.	**Supporting information**				
		E.g. external documents	2.8	Reference to additional files of relevance	

S1.5 References

Aschner, M., Ceccatelli, S., Daneshian, M. et al. (2017). Reference compounds for alternative test methods to indicate developmental neurotoxicity (DNT) potential of chemicals: Example lists and criteria for their selection and use. *ALTEX 34*, 49-74. doi:10.14573/altex.1604201

Bal-Price, A., Hogberg, H. T., Crofton, K. M. et al. (2018). Recommendation on test readiness criteria for new approach methods in toxicology: Exemplified for developmental neurotoxicity. *ALTEX 35*, 306-352. doi:10.14573/altex.1712081

Flood, S., Houck, K. and Grulke, C. (2017). Development of a context-rich database of ToxCast assay annotations. doi:10.23645/epacomptox.5178610.v1

Hartung, T. (2007). Food for thought ... On validation. *ALTEX 24*, 67-80. doi:10.14573/altex.2007.2.67

Hartung, T., De Vries, R., Hoffmann, S. et al. (2019). Toward good in vitro reporting standards. *ALTEX 36*, 3-17. doi:10.14573/altex.1812191

Kinsner-Ovaskainen, A., Rzepka, R., Rudowski, R. et al. (2009). Acutoxbase, an innovative database for in vitro acute toxicity studies. *Toxicol In Vitro 23*, 476-485. doi:10.1016/j.tiv.2008.12.019

Leist, M., Efremova, L. and Karreman, C. (2010). Food for thought ... Considerations and guidelines for basic test method descriptions in toxicology. *ALTEX 27*, 309-317. doi:10.14573/altex.2010.4.309

Leist, M., Hasiwa, N., Daneshian, M. et al. (2012). Validation and quality control of replacement alternatives – Current status and future challenges. *Toxicol Res 1*, 8-22. doi:10.1039/c2tx20011b

Leist, M. and Hengstler, J. G. (2018). Essential components of methods papers. *ALTEX 35*, 429-432. doi:10.14573/altex.1807031

Leung, V., Rousseau-Blass, F., Beauchamp, G. et al. (2018). ARRIVE has not ARRIVEd: Support for the ARRIVE (animal research: reporting of in vivo experiments) guidelines does not improve the reporting quality of papers in animal welfare, analgesia or anesthesia. *PLoS One 13*, e0197882. doi:10.1371/journal.pone.0197882

Linge, J. P. and Hartung, T. (2007). ECVAM’s approach to intellectual property rights in the validation of alternative methods. *Altern Lab Anim 35*, 441-446. doi:10.1177/026119290703500411

Marx, U., Andersson, T. B., Bahinski, A. et al. (2016). Biology-inspired microphysiological system approaches to solve the prediction dilemma of substance testing. *ALTEX 33*, 272-321. doi:10.14573/altex.1603161

OECD (2017). Guidance Document for Describing Non-Guideline In Vitro Test Methods. *Series on Testing and Assessment No. 211*. OECD Publishing, Paris. doi:10.1787/9789264274730-en

OECD (2018). *Guidance Document on Good In Vitro Method Practices (GIVIMP)*. OECD Publishing, Paris. doi:10.1787/9789264304796-1-en

Pamies, D., Bal-Price, A., Chesne, C. et al. (2018). Advanced good cell culture practice for human primary, stem cell-derived and organoid models as well as microphysiological systems. *ALTEX 35*, 353-378. doi:10.14573/altex.1710081

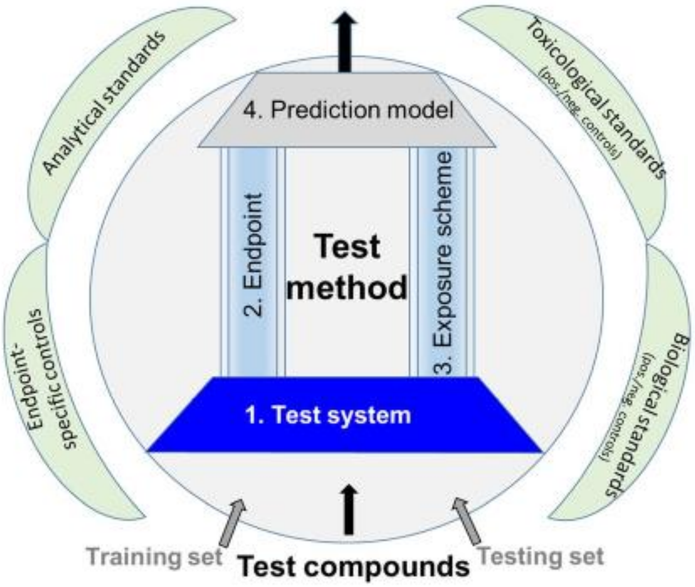
Prior, H., Casey, W., Kimber, I. et al. (2019). Reflections on the progress towards non-animal methods for acute toxicity testing of chemicals. *Regul Toxicol Pharmacol 102*, 30-33. doi:10.1016/j.yrtph.2018.12.008

Roi, A. J. and Flego, M. (2006). ECVAM’s DataBase service on alternative methods (DB-ALM) online. *ALTEX 23 Spec Iss*, 177.

Rovida, C., Vivier, M., Garthoff, B. et al. (2014). ESNATS conference – The use of human embryonic stem cells for novel toxicity testing approaches. *Altern Lab Anim 42*, 97-113. doi:10.1177/026119291404200203

Schmidt, B. Z., Lehmann, M., Gutbier, S. et al. (2017). In vitro acute and developmental neurotoxicity screening: An overview of cellular platforms and high-throughput technical possibilities. *Arch Toxicol 91*, 1-33. doi:10.1007/s00204-016-1805-9

S1.6 Appendix: Background and glossary



Test (or test method)

This term is used in many disciplines, and it is little defined in colloquial language. In toxicology, it is the term used to describe a procedure based on a test system, used to obtain information on the potentially hazardous effects of a substance. A toxicological test method consists of four major elements (i.e., 1. test system, 2. endpoint, 3. exposure scheme, 4. prediction model), and it produces a test result (information regarding the ability of a substance or agent to produce a specified biological/toxicological effect under specified conditions). Note that besides these four technical elements, the test purpose can be regarded as a fifth element. The term "test method" is used interchangeably with "test" and "assay" in the literature. A test method can have several analytical endpoints, which can be combined to various outcome measures.

Test system (biological system)

This term is often confused with "test method", but it has a different definition. A test system is a cellular (or biochemical) system used in a study (e.g., "proliferating neural stem cells", "neuronally-differentiating PC-12 cells", or "organotypic hippocampal slices"). The term is often used interchangeably with "*in vitro* system". The term test system is equivalent to "biological model" as far as test setup is concerned. From this follows that the test system is only one component of a test or "test method". Good performance of a test system does not imply good functioning of a test method. Acceptability criteria for test systems (e.g., at least 75% of the differentiated cells staining positive for nestin under control conditions) are different from acceptability criteria for the test method using the test system (e.g., inhibition of differentiation by a specified positive control by at least 35% and alteration of normal differentiation by a defined negative control by less than 10%).

Endpoint / test endpoint

The term endpoint has two implications, and it is essential to understand the differences. Within the context of a toxicological test, the endpoint is the biological or chemical process, response, or effect assessed in a test system by a specific analytical method/assay. For instance, "cell viability", "cell proliferation", or "electrical network activity" are endpoints. Each endpoint may be assessed by different "analytical methods". For instance, "viability" may be assessed by LDH-release, resazurin reduction, cell counting or measurement of ATP. "Differentiation" may be measured by PCR quantification of a differentiation marker or by morphometry (e.g., dendritic tree arborizations or synaptic spine density).

Analytical endpoint

An endpoint of a test system (e.g., proliferation, differentiation, or viability) may be quantified by different analytical methods (measurement endpoints). It is important to distinguish such analytical endpoints (referring to the analytical methods used) from (test system) endpoints that refer to the biological concept being evaluated. The test endpoint and analytical endpoints require independent optimization, characterization, and use of control compounds.

Exposure scheme

A drug may be added to a test system continuously or for certain time periods, in a certain solvent, with or without medium change, at a specified temperature, etc. All this information is contained in the exposure scheme. As each of the other three elements of a test, an exposure scheme needs to be optimized independently. For instance, with all other test parameters fixed, the test outcome can dramatically change with the time period of exposure. Depending on the point-of-view, the analytical endpoint may be regarded as part of the exposure scheme. Optimization of the exposure scheme may require switching analytical endpoints, even if the same test endpoint is evaluated.

Prediction model

The prediction model (PM) is a formula or algorithm (e.g., formula, rule, or set of rules) used to convert the results generated by a test method into a prediction of the (toxic) effect of interest. Also referred to as decision criteria. A prediction model contains four elements: (1) a definition of the specific purpose(s) for which the test method is to be used; (2) specifications of all possible results that may be obtained, (3) an algorithm that converts each study result into a prediction of the (toxic) effect of interest, and (4) specifications as to the accuracy of the prediction model (e.g., sensitivity, specificity, and false positive and false negative rates). The PM is often neglected in test setup. In its narrow sense, it defines the procedure how data are processed and how technical data (instrument readings) are translated into toxicological information. For instance, if calcium oscillations are measured, the PM determines what type of change is considered relevant to toxicity. Another important example is a change of gene expression, measured by PCR or a transcriptomics approach. A heatmap of gene expression is a technical set of data, but not toxicological information. A PM transforms this into a test statement of compound hazard. A first consideration about PM is whether there is a binary outcome (toxic/nontoxic) or are there more than 2 classes (mild, moderate, severe irritants), and how the boundaries are defined. For instance, many *in vitro* tests give information on whether a compound is hazardous or non-hazardous, but not on the strength of effect or the potency of a chemical. Another important issue is: If there are two or more assay endpoints (e.g., viability and neurite growth), how are they combined to a final toxicity statement? During test optimization and validation, the prediction model needs scrutiny and the questions asked are as follows: Is there a threshold (different from the statistical threshold) for when an effect can be considered biologically relevant? How is the outcome interpreted when more than one endpoint is measured (e.g., general cytotoxicity and functional impairment or effects on two different cell types)? Is an increase compared to normal good when a decrease is bad? How should data be interpreted when a compound alters the baseline values for the endpoint (e.g., a colored compound in a spectrophotometric assay)? What is the correct reference value if the test system changes over time? The PM defines these decision points and then translates the test result into a prediction, e.g., converting the luminometer reading of an ATP assay into a toxicological statement (prediction) on whether the compound is cytotoxic (at a given concentration). In practical terms, a test is set up to be predictive for unknown compounds (test compounds), but to achieve this goal, the different elements of the test usually require optimization and fine-tuning. This is performed by anchoring the test or its elements to a frame of known information, i.e., defined controls and standards.

Test purpose

Any test (toxicological or not) is developed to probe a test hypothesis (e.g., is a substance toxic or not). The test design will always reflect that purpose, and test parameters will (ideally) be optimized in order to achieve maximum certainty about whether the hypothesis should be accepted or rejected. It is a basic scientific principle that test results should – within limits – only be used for the purpose they were designed for. This is definitely not trivial for *in vitro*/NAM test methods developed for regulatory purposes. In this domain, the fifth test element (test purpose) plays a special role: Beyond the primary purpose (e.g., determination of cytotoxicity), the results of a test may also be used for a secondary regulatory purpose (regulation), and a third purpose (e.g., modelling a potential hazard in the population).

Analytical standards

Each analytical method requires calibration by the use of standards (positive and negative controls). This can include physicochemical approaches (e.g., to make sure that the balances and the spectrophotometer are working), or scaling approaches (e.g., to obtain absolute values in microscopic morphometric measurements or counts). On the next level, the analytical endpoint needs to be calibrated in the context of the test system. For instance, if LDH-release is used as a measure of viability, then it needs to be evaluated how much LDH is released under conditions of all cells dying (e.g., detergent lysis; not necessarily = 100%), and the overall assay needs to be normalized to such values. An important example is viability measurement by resazurin or tetrazolium dye reduction. This works only after normalization for cells that are 100% dead or alive, as the instrument readings as such have no dimension.

Endpoint-specific controls or mechanistic control compounds (MCC)

Chemicals known to reliably and consistently alter the endpoint of a test system at a mechanistic level. These are also referred to as “endpoint-selective controls” or “mechanistic tool compounds” or “technical controls”. This would be the first set of compounds used during test system setup to obtain information on the biological/toxicological behavior of the test system and its dynamic range. Such control compounds can be used to define acceptance criteria.

Positive/negative control (PC/NC) or toxicological standards

An NC for a “test method” is a compound or condition that should not trigger a response, i.e., it should not change the endpoint from baseline. A PC is a compound or condition that triggers a response, i.e., a change of the endpoint from baseline in the right direction and to a certain specified extent (for more detailed descriptions see 5.5-5.8)

Acceptance criteria

Criteria defined before performing an assay to determine whether it is “valid”, i.e., whether the data can be used. Typical issues of acceptance criteria comprise: “Has the actual run or plate of the test method functioned (e.g., are the endpoint values for MCC and NC in the right range)”, “Is the test method performing within the desired range of variability (e.g., are the standard deviations of MCC/PC and NC in the right range)”. Note that acceptance criteria can (and should) also be defined for an “analytical endpoint” or for a “test system”. For instance, for a test system, the acceptance criteria may say that it is only valid if at least 400 cells were in the region of interest, or if at least 80% neurons were present in mixed cultures, or if the average neurite length was at least 4 cell diameters. Such test system acceptance criteria are not at all related to those used for the test method. In this context, it is important to rationalize that endpoints that are meaningful for the description of the biological system/test system may not be useful for the test method and *vice versa*. For instance, a person's body weight can be measured well on scales (to give a good readout on general growth characteristics of a person = biological system), but this endpoint will hardly respond to acute poisoning of the person. Instead, blood pressure or vomiting activity may be good measures of human poisoning (toxicological test), but they in turn give little information on the growth activity over time. In a neurotoxicity test for network activity, the extent of synaptic staining may be a good acceptability criterion for the test system, but it will not react to a glutamate receptor agonist; on the other hand, electrical activity pattern will be a very sensitive measure for glutamate receptor-affecting toxicants, but the synapse number will not change (upon acute exposure). Once the first three elements of the test system have been established, optimized and assembled to a test, the prediction model can be established to complete the test system setup. One standard procedure is to use a training set of chemicals, i.e., known positive and negative controls, and run them through the test. Based on the test data, a prediction model would be established that best suits the known information about which of the compounds should test positive or negative. In a second round of testing, a test set of compounds would be used (i.e., a new set of positive and negative controls). The data of these substances would be run through the prediction model to determine accuracy, specificity and sensitivity of the test system. Possibly, further adaptations would then follow.

Training set chemicals

This set should include chemicals known (preferably from *in vitro* systems) to reliably elicit a response, or no response, with respect to the endpoint of interest. The goal of using this set is proof-of-concept that the test method can rapidly and efficiently screen moderate numbers of chemicals with reasonable predictivity. A training set of chemicals can be used to optimize an assay (test method), to set acceptability criteria, and to build a prediction model.

Testing set chemicals

This set would be used to validate and possibly improve the prediction model. The goal of using "testing set chemicals" is also to demonstrate the ability to test larger numbers of chemicals.