Concept Article

Limitations and Uncertainties of Acute Fish Toxicity Assessments Can Be Reduced Using Alternative Methods

Martin Paparella1, *, Stefan Scholz2,*, Scott Belanger3, Thomas Braunbeck4, Pascal Bicheref5, Kristin Connors1, Christopher Faßbender6, Marlies Halder7, Adam Lillicrap8, Roman Liska9, Kristin Schirmer5,10,11, Gilly Stoddart1, Paul Thomas5, Susanne Walter-Rohde1,2

1Division of Medical Biochemistry, Biocenter, Medical University of Innsbruck, Innsbruck, Austria; 2UFZ – Helmholtz Centre for Environmental Research, Department of Bioanalytical Ecotoxicology, Leipzig, Germany; 3The Procter and Gamble Company, Cincinnati, OH, USA; 4Aquatic Ecology & Toxicology, University of Heidelberg, INF 504, Heidelberg, Germany; 5Knowledge & Research in Environment and Toxicology in Silico (KREATIS), L’Isle d’Abeau, France; 8PETA International Science Consortium Ltd., London, United Kingdom; 6European Commission, Joint Research Centre, Ispra, Italy; 7Norwegian Institute for Water Research (NIVA), Oslo, Norway; 8Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland; 9EPF Lausanne, School of Architecture, Civil and Environmental Engineering, Lausanne, Switzerland; 10ETH Zürich, Department of Environmental Systems Science, Zürich, Switzerland; 11German Environment Agency (UBA), Dessau-Roßlau, Germany

Abstract

Information about acute fish toxicity is routinely required in many jurisdictions for environmental risk assessment of chemicals. This information is typically obtained using a 96-hour juvenile fish test for lethality according to OECD test guideline (TG) 203 or equivalent regional guidelines. However, TG 203 has never been validated using criteria currently required for new test methods including alternative methods. Characterization of the practicality and validity of TG 203 is important to provide a benchmark for alternative methods. This contribution systematically summarizes the available knowledge about limitations and uncertainties of TG 203, based on methodological, statistical, and biological considerations. Uncertainties stem from the historic flexibility (e.g. use of a broad range of species) and constraints of the basic test design (e.g., no replication). Other sources of uncertainty arise from environmental safety extrapolation based on TG 203 data. Environmental extrapolation models, combined with data from alternative methods, including mechanistic indicators of toxicity, may provide at least the same level of environmental protection. Yet, most importantly the 3R advantages of alternative methods allow a better standardization, characterization and an improved basic study design. This can enhance data reliability and thus facilitate the comparison of chemical toxicity, as well as the environmental classifications and prediction of no-effect concentrations of chemicals. Combined with the 3R gains and the potential for higher throughput, a reliable assessment of more chemicals can be achieved, leading to improved environmental protection.

1 Introduction

1.1 The current use of the in vivo acute fish toxicity test

The assessment of fish toxicity is an integral part of environmental hazard and risk assessment of many regulations worldwide (OECD, 2012a; Scholz et al., 2013). One of the frequently used vertebrate animal tests for aquatic toxicity assessments is the Acute Fish Toxicity Test (AFT)1, which is typically conducted according to OECD Test Guideline 203 (TG 203) or similar

Disclaimer: The scientific views presented in this paper are those of the authors alone and do not necessarily reflect official views of their respective institutions.

Received June 5, 2020; Accepted September 16, 2020; Epub September 16, 2020; © The Authors, 2020.

ALTEX 37(8), Published online: doi:10.14573/altex.2006051

Correspondence: Dr. Martin Paparella, Division of Medical Biochemistry, Biocenter, Medical University of Innsbruck, Innsbruck, Austria (martin.paparella@i-med.ac.at)

For example in Europe between 2015 and 2017, AFT testing caused about 45% to 60% of animal use from aquatic species for eco-toxicity testing, i.e. about 34,000 to 50,000 of 71,000 to 84,000 fish and amphibians, with amphibians contributing about 0.5% (European Commission 2020). Report on the statistics on the number of animals used for experimental and other scientific purposes in the member states of the European Union in 2015-2017. Commission staff working document. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1581689520921&uri=CELEX:52020SC0010, Part 2/5, tables 9.2 on pages 19, 53, 88)
guidelines (OECD, 2019b, 2012a; US EPA, 2016; ISO, 1996a, b, c). The AFT is used for the prospective assessment of individual chemicals, particularly to derive, depending on local or international regulations, an environmental classification, a predicted no-effect-concentration (PNEC) and/or one potential element of the toxicity criterion for PBT (persistence, bioaccumulation, toxicity) assessment (ECHA, 2017a). Furthermore, in some countries, the AFT is also conducted for effluent testing (Norberg-King et al., 2018; Scholz et al., 2013) or to inform the use of the test concentration for the fish bioconcentration test or as range finders for many other tests with more specific endpoints (OECD, 2012b). The AFT is based on a 96-h acute exposure of juvenile fish, identified as such by length. The percentage of lethality observed at each concentration is used to calculate an LC50 (lethal concentration at which 50% of the animals die). For the standard full-concentration-response test, at least 5 concentrations with a minimum of 7 fish per concentrations are used, without replication, resulting in a minimum number of 42 animals per test compound or sample (OECD, 2019b).

According to an OECD review (OECD, 2012a), the earlier TG 203 design from 1992 lacked critical specifications for several experimental parameters including the test duration (it was “preferably” 96 hours), use of solvents and solvent controls, application of statistical methods, measurement of fish length and selection of a test species or multiple species from the numerous species recommended (Table S1/s/section 9). Consequently, in 2019, TG 203 was revised to include more specifications and some test adaptations, including (as far as possible) the need for a validated analytical method to document actual test concentrations. The update also refers to revised guidance on the appropriate use of solvents (OECD, 2019a). However, the revision has not been broadly changed the basic test design. For example, lethality was not replaced by moribundity as the definitive endpoint, the number of fish used per concentration remained 7 as a minimum and the number of recommended fish species increased further to 11 (Table S1/s/section 9).

Moreover, due to the limited stringency of some specifications in the early test protocols, which also partially apply to the revised version (fish species/strain/age cohorts, water conditions, use of moribundity or lethality, section 2.2.), the available data are very heterogeneous (section 2.4.1 and Table S1/s/section 9.1.; (Braunbeck et al., 2020)).

1.2 The current use of alternatives to the in vivo acute fish toxicity test

Two experimental alternative methods have been standardized, validated and included into the OECD Test guidelines programme: The Fish Embryo Acute Toxicity Test (FET), has been adopted as OECD TG 236 (OECD, 2013, 2011a, 2012c; Busquet et al., 2014). The Fish Gill Cell Line Acute Toxicity Test using the rainbow trout (Oncorhynchus mykiss) RTgill-W1 cell line has been scientifically validated (Fischer et al., 2019; Tanneberger et al., 2013; Natsch et al., 2018; ISO, 2019) and, in 2019 was included on the OECD WNT workplan as Project 2.63 for the development of a regulatory OECD test guideline (Table S1/s/sections 8.2, 9.2, 10.2).

Furthermore, computational approaches are available to predict acute fish toxicity either as freeware, such as US EPA TEST4, VEGA (Benfenati et al., 2013), or as commercial software, such as CATALOGIC® and iSafeRat5 (Thomas et al., 2019). Moreover, similarity of chemical structures and/or in vitro data may be used to form chemical categories with the purpose of supporting read across of existing experimental in vivo data within those categories. The OECD QSAR Toolbox6 may support such assessments (OECD, 2007, Low et al., 2013). Work towards more automated “big data” approaches is also in progress (Helman et al., 2019; Luechtfeld et al., 2018).

These alternative methods can provide mechanistic indicators for sub-lethal toxicity:

- Endpoints in the FET (coagulation of fertilized eggs, lack of somite formation, lack of detachment of the tail-bud from the yolk sac and lack of heartbeat) are mechanistic in that they provide more information than simply whether a fish is dead or alive. If needed, these endpoints could also be expanded to include other endpoints such as neurotoxicity (Stengel et al., 2018; Zindler et al., 2019; Klüver et al., 2015).

- Based on the hypothesis that acute fish toxicity is often due to nonspecific modes of action, endpoints such as metabolic activity and cell- and lysosomal membrane integrity measured in the RTgill W1 cell-line test (Fischer et al., 2019) could be considered as a mechanistic key event, even without a fully characterized Adverse Outcome Pathway (Volz et al., 2011).7

---

2 doi:10.14573/altex.2006051s
3 Abbreviations: AFT = Acute Fish Toxicity; CTD = Chemical Toxicity Distribution; CV = Coefficient of Variation; EC50 = Effective Concentration 50%; EcoTTC = Ecological Threshold of Concern; FET = Fish Embryo Toxicity Test; GHS = Globally Harmonized System; GIVIMP = Good In Vitro Methods Practice; IATA = Integrated Approach to Testing and Assessment; ICE = Interspecies Correlation-Estimation; ITS = Integrated Testing Strategy; LC50 = Lethal Concentration 50%; MAD = Mutual Acceptance of Data; MechAsA = Mechanism of Action; MoA = Mode of Action; NOAEC = No Observed Adverse Effect Concentration; PBT = Persistent Bioaccumulating Toxic; PNEC = Predicted No-Effect Concentration; QAR = Quantitative Activity-Activity Relationship; QSAR = Quantitative Structure-Activity Relationship; SSD = Species Sensitivity Distribution; TG = Test Guideline; TG 203 = OECD Test Guideline 203; TG 236 = OECD Test Guideline 236; VMG Eco = (OECD) Validation Management Group Ecotoxicology; WNT = Working Group of National Coordinators of the OECD Test Guidelines Programme; WoE = Weight-of-Evidence.
9 The key relevance of cytotoxicity as a mechanism for acute toxicity was also recognized for acute mammalian toxicity (Priet et al., 2019; Vinken and Blaauwber, 2017).
Since 2005, only one test using fish embryos, the fish-egg test (ISO, 2016) used in Germany as part of the waste water dues law, has been implemented as a stand-alone replacement of an acute fish toxicity test (Bundesgesetzblatt, 2005; Norberg-King et al., 2018). In contrast, in chemical regulation worldwide, none of the experimental or computational alternative methods have been fully accepted as a stand-alone replacement for TG 203. Some, such as the FET and computational approaches, are considered useful at least within a Weight-of-Evidence (WoE) approach (ECHA, 2017b). However, the WoE and read across-based approaches, which combine multiple sources of information, may be of limited regulatory efficiency and are rarely used due to their high complexity and low standardization. They could lead to subjectivity of data selection and integration, possibly resulting in disagreement between experts and low assessment throughput. Therefore, work towards a quantitative WoE approach to replace TG 203 using Bayesian networks has been initiated (Lillicrap et al., 2020; Moe et al., 2020).

A testing strategy which does not replace the TG 203, but reduces the number of fish required, is the threshold approach for acute fish toxicity, and this was standardized at OECD level (OECD, 2010). In this approach, standard acute toxicity tests not involving the use of vertebrate animals are first conducted with daphnids (OECD, 2004) and algae (OECD, 2011b). Using appropriate negative controls, fish are then exposed to the lowest EC50 of these tests at a single concentration (the threshold concentration) or using a limit test (100 mg/L), whichever concentration is lower. A full TG 203 concentration-response test is only performed if toxicity is observed at the threshold concentration. Since daphnids and algae are frequently the most sensitive trophic levels and, therefore, drive environmental classifications and PNECs, the threshold approach results in a significant reduction in the number of fish required for regulatory purposes (Jeram et al., 2005; Hutchinson et al., 2003). Recently, the possibility of including the FET in the threshold approach was explored (Rawlings et al., 2019) to support a time and cost-efficient use of the new method, and to optimize 3Rs gains for predicting acute fish toxicity.

The development of an “Integrated Approach to Testing and Assessment (IATA) for acute fish toxicity” was included in the OECD WNT work plan in 2015 (WNT project 2.54). In principle, an IATA for acute fish toxicity might be constructed similarly to the IATAs for skin or eye irritation (OECD, 2014, 2018a): First, a WoE assessment of all available and relevant information is conducted. This may already lead to a conclusion or could inform the need for follow-up testing using an Integrated Testing Strategy (ITS). The ITS may aim at estimating whether the LC50 for fish is lower than for daphnids and/or algae, and only if this is the case, a more in-depth estimate for the fish LC50 should be provided. Such an ITS could represent an alternative to the current threshold approach by starting with acute tests with algae and daphnids, followed by QSARs, fish cell lines, and/or fish embryos, and conditionally—and only as a last resort if indicated by the available data—would TG 203 be conducted. Computational approaches for data integration, e.g. Bayesian networks (Lillicrap et al., 2020; Moe et al., 2020), could complement the IATA, remove subjectivity and provide an output in terms of a probability for a result.

The development of such an IATA with low potential for ambiguity (see Tab. 1, last line) may be essential for practical regulatory use and predictable acceptance by all stakeholders.

### 1.3 Transparency of scientific uncertainty is essential for responsible decision-making

This argument was already provided elsewhere (Paparella et al., 2020), but it is repeated and adapted here for the specific context.

It is essential for responsible decision-making in the management of chemicals that the uncertainties in data and knowledge are transparently described. This is important, as risk assessment and decision-making are typically carried out by different regulatory units or bodies. Guidance and tools have been developed for transparent characterization of the uncertainty of chemical risk metrics, such as ratios between human exposure and human limit values (EFSA Scientific Committee et al., 2018; WHO, 2018).

However, there is a need for a similarly transparent analysis of uncertainties of the performance metrics for testing methods within the validation process. This has recently gained recognition in the field of human health regulatory toxicology, where, specifically, the uncertainty characterization of standard in vivo reference methods is starting to promote the acceptance of alternative methods. This experience from human health toxicology could also support the acceptance of alternative approaches to acute fish toxicity testing and assessment.

As an example, information on the reproducibility of test guidelines for animal tests in the field of eye-irritation/damage and skin sensitization sets limits for achievable correlations between data from alternative methods and the animal-test-based reference methods (Adriaens et al., 2014; Barroso et al., 2017; Hoffmann et al., 2018). It was also analyzed how the experimental variability of acute rodent LD50 data translates into variability of GHS classification (Hoffmann et al., 2010). Later it was highlighted that, from a scientific perspective, a borderline range between GHS potency categories should be established. Test results falling into this borderline range should be considered as uncertain due to limited reliability of any test result (Leontaridou et al., 2017; Dimitrov et al., 2016). A comparable finding regarding aquatic acute toxicity classification has already been identified by Rawlings et al. (2019).

A systematic summary of uncertainties of animal reference methods is useful also in cases where a fully quantitative uncertainty characterization is not possible, due to a lack of data and/or the complexity thereof. It allows at least a semi-quantitative and qualitative comparison of the performance and uncertainties of both in vivo and alternative approaches. This could support a best-informed decision on the acceptability of the new methodology. Such work was conducted for the rodent based carcinogenicity assessment (Paparella et al., 2017) and it is ongoing in the field of rodent based developmental neurotoxicity assessment (Paparella et al., 2020).
Recently, the OECD Validation Management Group Ecotoxicology (VMG Eco) recommended that the uncertainties associated with the OECD TG 203 in vivo acute fish toxicity tests should be compiled (2018, unpublished recommendations for updates of the fish testing framework (OECD, 2012a)). There are already several studies that analyze TG 203 LC₅₀ variability ([Hrovat et al., 2009; Scholz et al., 2016; Belanger et al., 2013; Busquet et al., 2014; Braunbeck et al., 2020); see section 2.2. and Table S1/sections 9.1 and 10). However, a more-in-depth summary of the potential limitations and uncertainties in variability and in environmental extrapolation of TG 203 is still lacking. The purpose of the present manuscript is to provide such a summary, applying an approach which has been used previously for the 2-year rodent cancer bioassay based carcinogenicity assessment (Paparella et al., 2017). This approach builds on the existing OECD IATA Guidance Document (OECD, 2016) and suggests using identical structures for the characterization of the current method and the alternative approaches, including their specific uncertainties. This approach will facilitate a comprehensive comparative assessment in qualitative and quantitative terms.

2 Limitations and uncertainties for the use of TG 203 – perspectives for reduction by alternative method-based IATAs

The use of the AFT, as conducted according to TG 203, is characterized by a number of limitations and uncertainties that could be reduced by using alternative methods, in the context of IATAs. For the presentation and discussion of these limitations and uncertainties a similar approach was taken as published earlier for regulatory developmental neurotoxicity (Paparella et al., 2020): For a top level overview, the main aspects of this discussion are illustrated in Figure 1 and further discussed in section 2. Table 1 explains the terminology applied.

![Fig. 1: Limitations and uncertainties for the use of the OECD TG 203 acute fish toxicity test versus alternative method-based IATAs](image_url)

Acute fish toxicity assessment based on TG 203

- **Limitations**
  - 3Rs conflict: vertebrate test, lethality endpoint
  - Low testing throughput
  - No mechanistic information

- **Uncertainties experimental variability**
  - Constraints of basic test design: ≥ 7 fish, 1 cohort, no tank replicate, no positive control, low LC₅₀ confidence
  - Flexibility of test design: 11 species, no. of fish, water conditions, varying metabolism, monobund vs. lethal

- **Uncertainties environmental extrapolation**
  - LC₅₀ / pragmatic assessment factors

**Alternative methods IATA**

- ↑3Rs & throughput for assessing more available & new, lower risk chemicals for ↑ overall environmental protection
- Mechanistic, indicators for (sub)lethal toxicity for ↑ overall environmental protection
- ↑Replicates, positive control, ↑ concentration range with ↓ 3Rs conflict for ↓ LC₅₀ uncertainty
- ↑Standardization & validation for ↓ LC₅₀ variability & uncertainty, ↑ global comparability of results
- Data-based extrapolation methods (ICE, SSD, CTS, EcoTTC) combined with alternative methods may provide at least the same level of environmental protection

↑ = increase, ↓ = decrease; CTD = Chemical Toxicity Distribution; EcoTTC = Ecological Threshold of Concern; ICE = Interspecies-Correlation-Estimation; SSD= Species-Sensitivity-Distribution. (Sources of images: 3R image from ALTEX; cell-culture/computer & pond-image composition free from pixabay.com; fish-image for mechanism free from OECD AOP homepage; black fish drawings from Stefan Scholz)
Tab. 1 - Terminology used within this manuscript

<table>
<thead>
<tr>
<th>Term</th>
<th>Explanation</th>
<th>Potential improvements by alternative methods within IATAs&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limitation</td>
<td>Known practical disadvantages, e.g. testing throughput</td>
<td>yes</td>
</tr>
<tr>
<td>Variability</td>
<td>Variability is due to (biology and/or test guideline caused) diversity. For example:</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td>Experimental variability is intrinsic to biology and due to flexibility in the TG. For a given TG, variability cannot be reduced with further knowledge.&lt;sup&gt;b&lt;/sup&gt;</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td>Environmental variability is intrinsic to biology and environment. It cannot be reduced with further knowledge.</td>
<td>no</td>
</tr>
<tr>
<td>Uncertainty</td>
<td>Uncertainty is due to limited knowledge about a true value including its (biology and/or test guideline caused) variability. For example:</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td>The LC&lt;sub&gt;50&lt;/sub&gt; confidence interval is broad, if the LC&lt;sub&gt;50&lt;/sub&gt; is close to the border of the concentration range tested and the concentration-response slope is flat.</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td>The absence of a validation study implies uncertainty (= limited knowledge) about robustness and experimental variability of a method.</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td>Awareness of the high diversity of real environments indicates uncertainty about quantitative knowledge of environmental variability.</td>
<td>no</td>
</tr>
<tr>
<td>Complexity</td>
<td>Complexity stems from multi-causal effect-relationships, e.g. in hazard characterization for aquatic life, based on a WoE assessment (including, e.g., read-across, QSAR, animal test data from superseded TGs, new alternative methods data), which is the result of a series of decisions, including, e.g. data sources, data quality assignments for their selection, similarity measures for read-across, weight assigned to various types of data in relation to mechanistic knowledge.</td>
<td>no</td>
</tr>
<tr>
<td>Ambiguity</td>
<td>Uncertainty stemming from the plurality of scientifically legitimate viewpoints, e.g. resulting from the complexity of scientific assessments.</td>
<td>no, by alternative methods, yes, by IATA guidance&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup> The table is adapted from an earlier ALTEX manuscript (Paparella et al., 2017); for concepts of variability and uncertainty see e.g. (EFSA Scientific Committee et al., 2018) and for concepts of complexity and ambiguity see e.g. (IRGC, 2017)

<sup>b</sup> For discussion, see Figure 1 and text in section 2 of this manuscript.

<sup>c</sup> Knowledge about variability may be used to improve/change the test design. However, this results in a new TG with its new variability.

<sup>d</sup> IATA guidance reduces the ambiguity by rules agreed a priori to testing and assessment. For example, the IATA for eye damage/irritation prescribes to carry out a Weight of Evidence (WoE) assessment based on available data a priori to new testing and the result of this WoE determines the use of either of three different sequences of in vitro tests, and also the results at each step within the sequence of tests determines the need for follow up testing (OECD, 2018a).

In the supplement Table S1<sup>2</sup>, information about the limitations and uncertainties of the use of TG 203 is presented within an OECD standard tabular format, which was originally developed to characterize alternative methods as individual information sources to be used within IATAs (OECD, 2016). This tabular summary was applied to carefully consider all potential limitations and uncertainties of the use of TG 203 and to develop the figure and text for section 2 in this manuscript. Table S1<sup>2</sup> may be further amended and refined, as far as useful, in the OECD Validation Management Group Ecotoxicology (VMG Eco). Applying the same systematic characterization scheme for both, the use of TG 203 and alternative methods may support regulatory decision-making on the acceptability of the latter. The table could also be included in alternative methods-based IATA guidance documents, as was the case for the standard in vivo study for the eye irritation/damage IATA (OECD, 2018a).

2.1 Limitations of TG 203 in terms of 3Rs, testing-throughput and mechanistic information – versus alternatives

Several limitations of TG 203 are inherent to the principle of this test and affect its practical regulatory applicability: this guideline is in direct conflict with the 3R goals (Russell and Burch, 1959), since it requires, for full concentration-response testing, at least 42 vertebrates of juvenile stages (a number causing statistical uncertainties, see section 2.2) and this does not take into account the need for a range-finding study. Moreover, according to the TG, lethality shall be used as an endpoint, which further aggravates the concern. Termination of acute toxicity testing when moribundity is observed would result in improved animal welfare (Russi, 2012), which is legally mandatory in Europe<sup>4</sup> and is currently applied in several countries. However, as explained below (section 2.2, paragraph 4), replacing lethality with moribundity as the endpoint was not agreed globally at OECD level in the last update of TG 203.

The assay can only be conducted in a low-throughput manner. Rearing of fish is required to obtain the suitable juvenile size, which may take several weeks, depending on the species. Testing as such requires 96 hours, not including the additional time for planning, preparation, assessment, reporting and tracking of culture health for approximately two weeks prior to testing. The throughput is further limited by the large volumes and vessels required to conduct the text. Therefore, it appears principally difficult to provide data for the more than 100,000 chemicals in commerce, for which toxicity data are currently lacking<sup>5</sup> (Table S1<sup>2</sup>/section 16).

<sup>4</sup>The estimate is based on the number of chemicals in the ECHA Classification and Labelling inventory, i.e. about 142,000 that should be on the European market. About 22,000 chemicals are registered for REACH in more than 1 ton per year, for which acute toxicity studies for daphnids and algae are required, but only substances registered above 10 tons per year have requirements for acute fish toxicity, these are about 7,000 chemicals. Slightly more than 60% of the data-requirements were
Moreover, the current endpoints in TG 203 provide little mechanistic information, which could be useful for read-across and inferring chronic toxicity, or supporting interspecies/environmental extrapolation modelling (cf. section 2.3, first two green shaded paragraphs). Introducing moribundity and eventually additional mechanistic endpoints by default into TG 203 would be theoretically feasible. However, the current situation of limited quantity and quality of ecotoxicity data should be improved, e.g. according to the European Strategy for a non-toxic environment (European Commission, 2017) and the US vision for toxicity testing in the 21st century (National Research Council, 2007) and reaching this goal with traditional animal testing seems impossible.

In contrast, alternative approaches allow an increase in the testing throughput by using small scale assessments and possibilities for automation. Alternative approaches would also allow testing of environmental degradation and reaction products (e.g. from disinfectants with biological material) and mixtures. Eventually, new chemicals that may have reduced environmental risk could be tested, when available in laboratory-scale amounts only and this may also promote the development of “green chemistry” (Maertens et al., 2014). In summary, by providing more data for assessing many more chemicals and mixtures, alternative approaches may contribute to an improved environmental safety without compromising global 3Rs goals (Table S1²/section 16.1).

2.2 Uncertainties in experimental variability relating to the study design of TG 203 – versus alternatives

The basic study design of TG 203 causes uncertainty. Given the variability between individual fish and the use of a minimum of 7 fish per concentration from one cohort without tank replicates may lead to broad confidence intervals in LC₅₀ estimates derived from concentration-response modelling, especially in the case of flat concentration-response relationships or when the LC₅₀ is off-center relative to the boundaries of the tested concentration range (Table S1²/section 5.1; (Carr et al., 2018)). In addition, while the absence of study-internal positive controls is important to prevent further animal use, this causes uncertainty about potential intra- and inter-laboratory variability (Table S1²/section 4.3.).

In contrast, alternative experimental methods allow for an improved basic study design in terms of replicates, concentration ranges and inclusion of study-internal positive controls. This is possible due to their small scale, potential for automation and 3Rs benefits. This improvement can reduce the uncertainty of results.

Some uncertainties of TG 203 relate to its level of standardization: for instance, TG 203 is flexible regarding the use of the test species. Any one of 11 recommended test species may be used, and guidance for selecting any of these is very generic. According to TG 203, species selection should depend “on regulatory requirements (industrial chemical, pharmaceutical, biocide or plant protection product, etc.) and on environmental exposure scenarios (cold, temperate or warm water species, freshwater or estuarine/marine fish)” (OECD, 2019b). The possible use of diverse fish-strains adds to this uncertainty. Other variables in the study design may also affect LC₅₀ estimates, such as the test species-related water conditions (temperature, salinity, water hardnes, pH). Also, the potentially variable age cohorts may affect the toxicity estimates (small differences within the recommended length range translate by cubic function to larger ranges of weight and developmental stage; cf. Table S1²/section 3 as well as Table 6 in Belanger et al. 2013). This diversity of potential test-designs also means variability and uncertainty in variability of biotransformation in the AFT, little is known about species differences in this regard (cf. Table S1²/section 8, (Braunbeck et al., 2020; Schlenk et al., 2008)).

Furthermore, LC₅₀ estimates in regulatory practice may be impacted by the inconsistent use of the endpoints lethality and moribundity (to conform with OECD TG 203 and Directive 2010/63/EU, respectively). On the one hand, the use of moribundity may reduce LC₅₀ estimates on average by a factor of 2 (Rufli, 2012). On the other hand, observations of moribundity are likely more subjective than observations of mortality and may introduce additional variability to the assay result. However, the variability and uncertainty from the use of moribundity or lethality can be estimated and reduced as soon as unambiguous criteria for moribundity have been agreed (Table S1²/sections 1.1. & 4.1.; (Rufli, 2012)). Thus, there is still uncertainty related to the use of these endpoints, but in principle the difference between lethality and moribundity could be scientifically calibrated (Table S1²/section 4.1.).

An assessment of AFT LC₅₀ values indicates a variability of up to a maximum range of 6 logarithmic units for the same chemical. However, this is based on historical data without application of stringent data quality filters (Hrovat et al., 2009). Two older ring-trials indicate a maximum range of LC₅₀ inter-laboratory variability of one logarithmic unit, if fish interspecies variability is excluded, but variability from other aspects such as fish size and exposure conditions (flow-through or static) is included. The two assessments were based on one chemical each, in one or two replicates, one or two fish species within 6 or 13 laboratories (Lemke, 1981; US EPA, 2001); Table S1²/section 10 also includes CVs for comparison with other ecotoxicity tests). However, a more comprehensive and most recent analysis applying stringent data quality filters indicated that about 8% and 0.5% of 181 chemicals showed differences in AFT LC₅₀ data by factors of >10 or >100, respectively, if interspecies variability is excluded. If the TG 203 inherent interspecies variability is included these percentages increase to about 15% and 10% of 53 chemicals ((Braunbeck et al. 2020; Table S1²/section 10). Also other work including AFT data for...
the neurotoxic biocide malathion indicates that AFT interspecies difference may be in the range of 4 orders of magnitude, depending on the chemical (Figure 5 in Fischer et al. 2019). For compounds that require bioactivation, differences in the LC50 of 50-500-fold have been identified for different fish species (Scholz et al., 2016).

Overall, it is uncertain how the combination of all the variables within the TG 203 study design impact on the LC50 value (Table S1/sections 9 & 10). In principle, TG 203 could be standardized more stringently, similarly to TG 236 and more recent in vitro methods. This might reduce variability and uncertainty in LC50 values. However, at the OECD level, more standardization of TG 203 was not intended during the recent update process.

As theoretically indicated in the TG, the current flexibility could favor lethality estimates for a specific fish species and its specific environment or accommodate regulatory preferences. Moreover, the flexibility favors the practicability. Yet, increasing standardization of TG 203 now would not resolve the current heterogeneity of the historical database (section 2.4.1) and could trigger a huge increase in regulatory demands for retesting with animals, which would conflict with current 3R goals. It would also not provide the desired significant increase in overall global environmental protection, since the throughput remains limited and the uncertainty related to the intention for more specific environmental extrapolations is underestimated (sections 2.3. and 2.4.2).

Nonetheless, compared to TG 203, the experimental protocols of alternative methods already provide a higher level of standardization. This may reduce the variability of test results, if the OECD guidance on good in vitro method practice (GIVIMP) was followed to develop robust protocols (OECD, 2018b), like those for the RTgill-W1 test (Table S1/sections 3.2, 10 and 10.2). Moreover, comprehensive validation studies are available for alternative methods with well-standardized test protocols. Therefore, the uncertainty in the experimental variability estimates of test results may be relatively low (Table S1/section 9.2.). Low variability and especially low uncertainty in variability are a significant advantage for environmental protection and global regulation in terms of PNEC calculation, GHS classification or the identification of the toxicity criterion for PBT assessment. Alternative methods may allow an increase in the global comparability of test results and thus reliably identify and globally regulate the – relative to all chemicals on the market – more toxic chemicals.

2.3 Uncertainties in environmental extrapolation – TG 203 versus alternatives

A relevant improvement for TG 203 would be to use moribundity instead of lethality as the endpoint. Lethality can represent a rather crude indicator for a chemical’s potential to cause a population decrease in real environments (which is the ultimate intention of ecotoxicity testing). Moribundity might be an environmentally more relevant and protective endpoint. This hypothesis appears mechanistically plausible, since a weakened fish is likely to impact on populations in complex environmental situations, which include predators, competitors and/or other environmental stressors (Zhao et al., 2020; Knillmann et al., 2012; Ruffi, 2012). However, as explained above (section 2.2. paragraph 4), this potential improvement has not been agreed globally at the OECD level yet.

TG 203 is used in an attempt to assess the acute toxicity of a chemical to fish – based on the estimation of toxicity observed in only one developmental stage (juvenile) of one test species. LC/EC50 data from different organisms (fish, invertebrates, algae) are used in combination with pragmatic assessment factors to account for the potential variability in the sensitivity of different aquatic trophic levels (ECHA, 2008). Yet, the aquatic environment contains hundreds of thousands of species (Mora et al., 2011), various life-stages and a vast array of abiotic and biotic modifiers. Data-based knowledge is available demonstrating that tests using a single species are “in a majority of cases, reliable qualitative (some level of response seen) predictors of aquatic ecosystem community effects” (de-Vlaming and Norberg-King, 1999). This latter US EPA review identified 57 studies (74 %) that support this conclusion, 16 studies (21 %) where single species testing underestimated aquatic ecosystem effects, and 4 studies (5 %) that were inconclusive. The review also explains that full quantitative validation of single species tests through field studies is neither feasible nor meaningful given the huge environmental variability. This also means that a single environmentally “true” value does not exist, regardless of the assessment method applied (see below, last paragraph in green section and Table S1/section 2.1). However, the standard assessment factor approach results in a PNEC with an unknown level of environmental protection, in terms of proportion of species under risk and related uncertainty. For GHS classification of chemicals, no assessment factors, but pragmatic cut-off values for the LC50 or EC50 from fish, daphnids and/or algae are used (as far as available), eventually in combination with information on biodegradability and bioconcentration. This represents a similarly pragmatic approach (Table S1/sections 6 & 7).

In summary, stand-alone LC50 values from TG 203 provide quite an uncertain basis for estimates of environmental toxicity (Table S1/section 2).

Endpoints tested within alternative methods are not intended to be specific for any fish species, but rather as a useful basis to estimate fish toxicity at least for all current standard species in TG 203. Moreover, the endpoints may represent mechanistic indicators of an increased probability of fish population level lethality, relative to other chemicals (section 1.2. and Table S1/section 1.2.). Their mechanistic information content may also support inferring potential long-term impact as well as interspecies/environmental extrapolation modelling (see next paragraph). Considering the environmental extrapolation uncertainties associated with TG 203-derived PNECs and GHS classifications, alternative approaches may be expected to provide at least a similarly (uncertain or even improved) environmental protection level. For example, differences between LC50 values derived from TG 203 and zebrafish embryo tests (e.g. OECD TG 236) have been found to be in the range of the fish species variability accepted within TG 203 (Lammer et al., 2009; Scholz et al., 2016).

Moreover, computational approaches are available for extrapolating experimental LC50 values to other environmental species. Interspecies Correlation Estimates (ICE), Species Sensitivity Distributions (SSDs), Chemical Toxicity Distributions (CTDs) and Ecological Threshold of Concern (EcoTTCs) models (Bejarano et al., 2017; Belanger et al., 2015; Connors et al., 2019) provide estimates for a predicted species or trophic level effect without the need for additional animal test data. It is noted that the data used for model development stem from laboratory experiments and can never inform on the almost endless real environmental variability, but they allow best informed use of available knowledge (Table S1/section 6.2). However, the
development of these models and variability of model output might also be improved by less variable data from alternative methods.

It may be argued that the use of alternative methods introduces some additional and new uncertainties. Yet, these additional uncertainties may not significantly increase the current uncertainties for environmental safety assessment based on TG 203. Moreover, they may not be conceptually different (see Table 2).

In summary, a combination of alternative methods would allow the assessment of a larger number of chemicals, thereby promoting the identification and new development of chemicals with lower environmental risk. The improved basic test design and better standardization of alternative methods may also increase our ability to compare the toxicity of chemicals. In combination with environmental extrapolation models, data from alternative methods may provide mechanistic indicators of toxicity with at least the same level of environmental protection as current approaches.

### Tab. 2: Potential additional uncertainties from the use of alternative methods compared to similar uncertainties from current TG 203 based environmental safety assessment.

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Alternative methods used for environmental safety assessment</th>
<th>In vivo AFT used for environmental safety assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novel chemical with unknown/ novel MOA</td>
<td>For completely new chemicals, which possess currently unknown modes of action, the relevance of the alternative methods results may be more uncertain than for known chemical domains/MoAs. This is, because such chemicals may not have been covered within the available environmental data analyses used for the generation and/or the performance assessment of alternative methods and computational models.</td>
<td>It is uncertain, how many of the MoAs most relevant for environmental safety are covered with the AFT: The chemical domain of applicability has not been formally defined for the AFT. It is unlikely that the AFT, which is not covering effects on embryonic development, covers all possible environmentally relevant MoAs. Bearing in mind the hundreds of thousands of species and specific life-stages present in the aquatic environment (Mora et al., 2011), it is also uncertain which MoAs may be relevant for acute aquatic toxicity, but are not covered by testing usually only one species for each of the three trophic levels (fish, daphnids and algae).</td>
</tr>
<tr>
<td>Relevance to organism and population level effects in the natural environment</td>
<td>Molecular/cellular level effects used as mechanistic indicators for potential toxicity may or may not be compensated at organism and/or population level. Such knowledge may remain limited due to the high complexity of mechanisms and interactions at organism, population and ecosystem level. (FET represents a test with an intact organism, thus uncertainty is anyway limited to the extrapolation from the laboratory test to the environmental population and ecosystem level)</td>
<td>Sublethal effects and lethal effects at organism level in the laboratory may translate to various responses at population level in the different ecosystems, depending on the specific fauna and flora present, the specific food-webs, competitors and/or manifold other environmental variables and potential stressors (Zhao et al., 2020). Acute toxicity effects in the laboratory are “just” indicators for potential real-world environmental effects; see also discussion above (de-Vlaming and Norberg-King, 1999).</td>
</tr>
<tr>
<td>Biotransformation</td>
<td>In vitro biotransformation is limited to the biotransformation capacity of the isolated test system and may be dissimilar to in vivo metabolism. Biotransformation in FET may differ from AFT.</td>
<td>Biotransformation may vary between species, life stages, and in response to environmental factors. Knowledge of this real-world environmental variability is very limited (Table S13/section 8).</td>
</tr>
<tr>
<td>Sensitivity across fish species</td>
<td>In vitro methods currently do not provide information on species-sensitivity differences. Theoretically FET may be carried out with different species. In vitro assays could also be created using cell lines from different fish species.</td>
<td>AFT can be conducted with a limited number of fish species that can be easily reared and/or tested in laboratory conditions. Furthermore, typically only one fish species is tested or required for regulation and hence, in regulatory practice no chemical-specific species sensitivity comparison is conducted. Such testing does not include the variability of fish toxicity due to variable factors in the real-world environment. Moreover, fish may not be the most sensitive aquatic species. Hence chemical specific data about fish-species variability may not significantly reduce uncertainty for environmental protection. (Table S13/sections 2.1. and 5 to 9).</td>
</tr>
</tbody>
</table>

2.4 Are further data needed to characterize the uncertainty of TG 203 data used in environmental hazard and risk assessment?

#### 2.4.1 Uncertainties in experimental variability

TG 203 has never been formally validated. LC50 variability is uncertain, both for identical study designs and for all the study design variants covered within TG 203 (section 2.2). It is not known, how the variables within the TG 203 study design
(selection of fish species, water conditions, life-stage, sex) affect the LC50. If we had a better knowledge of how this variability in the test design can affect LC50 values, we would be able to calibrate any specific data set. Several reliability estimates for acute fish toxicity data have been published, and these indicate a concern (section 2.2, Table S13/section 10). In order to better describe assay variability, a broader and more carefully curated AFT database could improve the quantitative assessment of variability (Braunbeck et al., 2020). However, such an extended retrospective assessment may be difficult or not feasible because robust historical data are generally scarce, i.e. about 15% of all available data (Braunbeck et al., 2020), and not available for all chemical groups and physical-chemical properties. Moreover, information on the variability of biotransformation in different fish species and life stages is scarce (Schlenk et al., 2008; Braunbeck et al., 2020). Therefore, it appears prudent to use the available data and other information to inform potential advances in regulatory science and decision-making for the acceptance of new, alternative approaches.

It should be considered at the regulatory science-level whether it would be useful to develop a system using probabilistic (instead of deterministic) assignment of chemicals to the acute aquatic GHS category and the associated M-factors13. Categorization and M-factor attribution could be expressed in terms of a probability that the LC50 value is higher or lower than the acute category cut-off value of 1 mg/L and that the LC50 value will be within any of the 10-fold M-factor stratifications. In case the current variability of the standard threshold approach based LC50 or EC50 (OECD, 2010) appears to be very high, refinement of the current GHS M-factor stratification might be considered. In view of the limitations, variability and uncertainties stemming from the use of TG 203 and the advantages of using alternatives, revision of the GHS classification criteria to explicitly include alternative approaches should be considered.

2.4.2 Uncertainties in environmental extrapolation

The need to use practically feasible approaches necessarily limits any type of testing and assessment in terms of its predictive capacity for the highly variable and complex aquatic environment. This, of course, also applies to the use of alternative approaches. It might be helpful to consider that any science-informed regulatory assessment needs to rely on some type of extrapolation model (Table S13/section 6) – and this need indicates a conceptual similarity of animal tests and alternative tests. Several tools, including ICE, SSD, CTD and EcoTTCs models can be used to describe, model, and account for the variability across species (Table S13/sections 6.1-6.2). On this basis a single experimental fish LC50 (or prediction thereof) can be recognized as a very uncertain estimate for environmental aquatic toxicity. Confidence intervals for more comprehensively informed environmental toxicity estimates, such as the 5th percentile of SSDs may span 2 orders of magnitude (Bejarano et al., 2017; Awkerman et al., 2014). However, the SSD models do not include the variability of fish toxicity due to the variability of environmental factors. Moreover, it is important to acknowledge variabilities and uncertainties and review them with respect to the final regulatory use of the data. It should be recognized that any hazard characterization with standardized methods essentially represents a hazard comparison relative to other chemicals. Since in any case extrapolation uncertainties are huge, other aspects of scientific validity, such as biological/mechanistic relevance and low variability and low uncertainty in variability, should receive increased attention. Alternative methods with reduced variability and uncertainty in variability (relative to the current TG 203) would at least better support the desired reliable hazard comparison between chemicals in terms of PNECs, GHS classifications and identification of the toxicity criterion in the context of PBT-assessment. It may be considered that alternative approach-based LC50 estimates could directly be used for environmental extrapolation model development (such as ICE, SSD, CTD, EcoTTC). Computational approaches for extrapolation to environmental relevance might even be improved by the use of less variable data from alternative methods.

In summary, more knowledge to reduce the extrapolation uncertainty from TG 203 to the environment may not necessarily be needed now for the immediate use and continued development of alternative methods.

2.5 Considering complexity for decision-making

Scientific data selection and knowledge integration is a complex task that may lead to a situation that different expert-groups may come to different, scientifically legitimate conclusions. For example, PNECs derived by expert risk assessors, on the basis of current standards, can vary by 3 orders of magnitude, with the largest contributor being the heterogeneous judgment of study quality (Hahn et al., 2014). This represents ambiguity as one form of uncertainty, and it regularly appears within discussions on the regulatory validation and official acceptance of new alternative approaches. Validation is usually built on available complex data and information.

In situations of ambiguity, there may be a tendency to stick to the traditional approach. Typically, the existing animal test-based approaches for chemical hazard characterization are considered as the “gold standard”. Given the experience and long-term use of the test, there is a high perceived confidence in the relevance of the result – albeit a robust validation of these tests is often lacking. In contrast, for alternative approaches, detailed validation studies have often been conducted, and the intra- and inter-laboratory variability is known, but perceived confidence is limited.

Imagine that an alternative approach represented the established standard assay and the acute fish toxicity test represented the new approach (Braunbeck et al. 2020): Are the available data for the AFT convincingly superior to the alternative in terms of uncertainty, variability and environmental extrapolation? Would the available data be sufficient to replace the alternative method by the AFT?

---

13According to GHS and CLP regulation (EC No 1272/2008), LC50 values higher than 1 mg/L do not lead to acute aquatic toxicity classification, whereas LC50 values ≤ 1mg/L lead to classification into category 1. No other acute categories are defined, but further LC50 stratification into orders of magnitude (1 · 0.1 mg/L, 0.1 · 0.001 mg/L, etc.) allows the attribution of M-factors to the classified chemicals and this supports better more accurate classification by accounting for toxicity further to dilution of each constituent. Also classification for chronic toxicity categories is possible based on acute fish LC50 data, depending on information for ready biodegradability and log Kow/BCF, ECHA (2017c). Guidance on the Application of the CLP Criteria (Version 5.0). https://echa.europa.eu/guidance-documents/guidance-on-CLP
In case of ambiguity, preference should be given to alternative approaches that can provide significant gains in terms of 3Rs, testing and assessment throughput, costs and improved reliability. All these aspects are critically important to support testing and comparing the toxicity of many more chemicals in order to reduce the overall toxicological burden in the environment.

3 Conclusions

The current estimation of acute fish toxicity based on TG 203 bears various scientific uncertainties and practical regulatory limitations for achieving the final goal of protection of the environment from hazardous chemicals. The limitations relate to conflicts with the 3Rs principles, low throughput and lack of mechanistic information. Uncertainties relate to experimental variability stemming from the basic study design and the study flexibility. Further uncertainties relate to the need for extrapolation to the highly variable environment.

Considering the interest of significantly improving the level of environmental protection, it is desirable that more reliable and comparable test data be generated and assessed for many more chemicals on the market as well as new chemicals intended to lower environmental risks. To achieve this aim, the future focus of regulatory toxicology needs to shift from individual Weight-of-Evidence based substance assessment towards development and harmonization of IATAs. These should be built on highly standardized alternative methods, supported by computational approaches\textsuperscript{14}. Such a strategy may be particularly important for lower tier studies, which use relatively simple animal test guidelines with crude endpoints such as acute fish toxicity.

Supplementary material

Table S1\textsuperscript{2}: Characterization of the acute fish toxicity test (OECD TG 203), according to the OECD template for individual information sources to be used within IATAs.

References


\textsuperscript{14}Establishing GLP like quality control systems for computerized data assessment will also become an important aspect and this is starting to be addressed at OECD level. Good Modelling Practice principles will support relevant model development and assessment EFSA (2014).


ECHA (2017b). Expert Workshop on the potential regulatory application of the Fish Embryo Acute Toxicity (FET) Test under REACH, CLP and the BPR, 3-4 May 2017 in Helsinki.


Rawlings, J. M., Belanger, S. E., Connors, K. A. et al. (2019). Fish embryo tests and acute fish toxicity tests are interchangeable in the application of the threshold approach. Environ Toxicol Chem 38, 671-681. doi:10.1002/etc.4351


Conflict of interest
The authors declare not to have any conflict of interest with this manuscript.

Acknowledgements
The work of Martin Paparella at the Medical University of Innsbruck and the publication costs are financed by the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology, Department V/5 – Chemicals Policy and Biocides.