**Learned Discourses: Timely Scientific Opinions**

**Timely Scientific Opinions**

**Intent.** The intent of Learned Discourses is to provide a forum for open discussion. These articles reflect the professional opinions of the authors regarding scientific issues. They do not represent SETAC positions or policies. And, although they are subject to editorial review for clarity, consistency, and brevity, these articles are not peer reviewed. The Learned Discourses date from 1996 in the North America SETAC News and, when that publication was replaced by the SETAC Globe, continued there through 2005. The continued success of Learned Discourses depends on our contributors. We encourage timely submissions that inform and stimulate discussion. We expect that many of the articles will address controversial topics and promise to give dissenting opinions a chance to be heard. This section is dedicated to the memory of Dr. Peter M Chapman, who founded Learned Discourses and served as Editor until he passed away in 2017.

**Rules.** All submissions must be succinct: no longer than 1000 words, no more than 6 references, and at most 1 table or figure. Reference format must follow the journal requirement found at www.setacjournals.org. Topics must fall within IEAM’s sphere of interest.

**Submissions.** All manuscripts should be sent via email as Word attachments to the IEAM Editorial Office (learned_discourses@setac.org).

---

**RECOMMENDATIONS FOR REDUCING THE USE OF FISH AND AMPHIBIANS IN ENDOCRINE-DISRUPTION TESTING OF BIOCIDES AND PLANT PROTECTION PRODUCTS IN EUROPE**

Laurent Lagadic,† Katrin Bender,† Natalie Burden,† Edward R Salinas,§ and Lennart Weltje§

†Bayer AG, Crop Science Division, Environmental Safety, Ecotoxicology, Monheim am Rhein, Germany
‡NC3Rs, London, United Kingdom
§BASF SE, Agricultural Solutions - Ecotoxicology, Limburgerhof, Germany

DOI: 10.1002/ieam.4156 © 2019 SETAC

Criteria for determining the endocrine-disrupting (ED) properties of biocide (European Union [EU] Regulation 2017/2100) and plant protection product (PPP) active substances (EU Regulation 2018/605) entered into force on 6 June and 10 November 2018, respectively. Accompanying the criteria is a guidance document (ECHA–EFSA–JRC 2018), which provides recommendations for evaluating the ED properties of chemicals for humans and nontarget organisms. It recommends selected tests to fulfill data sufficiency for either endocrine activity or adverse effects (defined by WHO/IPCS 2009 as a change in the morphology, physiology, growth, development, reproduction, or life span of an organism, system, population, or subpopulation that results in an impairment of functional capacity, an impairment of the capacity to compensate for additional stress, or an increase in susceptibility to other influences). A test indicating endocrine activity usually triggers a test for adversity. The present article focuses on data needs using fish and amphibians.

The suggested tests for mechanistic data sufficiency are the Amphibian Metamorphosis Assay (AMA; OECD TG 231, 2009) and Fish Short-Term Reproduction Assay (FSTRA; OECD TG 229, 2012), whereas for adversity, the Larval Amphibian Growth and Development Assay (LAGDA; OECD TG 241, 2015b) and Medaka Extended One-Generation Reproduction Test (MEOGRT; OECD TG 240, 2015a) are recommended. These tests require the use of many animals, are highly complex, and are infrequently conducted. Tests that are infrequently conducted (e.g., OECD TGs 240 and 241) are at greater risk of “failure” as a result of deviation from test guideline validity criteria (Salinas and Weltje 2018). A test that fails validity criteria usually requires repetition,
doubling the number of animals used. Additionally, further animals are needed for range-finding tests to set the appropriate concentration range for the definitive testing.

Due to the required test organism numbers, this testing conflicts with the aim to reduce vertebrate testing and to employ vertebrate testing only as a last resort (see Regulations EC 1107/2009 and EU 528/2012 on PPPs and biocides, respectively). In the present article, we estimate the number of fish and amphibians required for ED testing, based on registered PPPs and biocides in Europe.

In order to estimate animal numbers, we first estimated the number of biocides and pesticides that require an evaluation for ED properties. The PPP active substances, safeners, and synergists with status “approved” and “pending” were extracted from the European pesticide database (EC, 2016). Excluded were attractants, desiccants, elicitors, plant activators, pruning substances, repellents, soil amendments, and viral inoculates. This extraction and exclusion process resulted in 475 PPP active substances. Biocides with status “approved” or “approval in progress” were selected from the European Chemicals Agency database (ECHA 2019). Duplication of substances was corrected for, yielding 297 biocide active substances.

The number of animals required for each test was estimated on the basis of information provided in the test guidelines (Table 1). We assumed and advocate for the use of 1 control in case a solvent is used. For the FSTRA, the fish species that required the smallest sample size (i.e., zebrafish) was selected as a conservative approach. For each test, a range-finding test was included with 3 test concentrations, a control and half the replicates of a definitive test. Also, a conservative failure rate (triggering test repetition) was estimated for each test based on the work of Burden et al. (2017), who surveyed test laboratories, and Salinas and Weltje (2018), who analyzed validation data.

Our estimates indicate that, for mechanistic data sufficiency, 304,000 animals would be needed to fulfill the requirement of ECHA–EFSA–JRC (2018) for pesticides and 190,080 animals for biocides. The estimates for MEOGRT and LAGDA are higher (1.9 and 1.2 million animals for pesticides and biocides, respectively). The use of such high numbers of animals is not compatible with the desire to reduce animal testing in the European Union (EU).

We propose the following ways to reduce the test animal numbers, without compromising the knowledge needed to conduct a proper ecotoxicity evaluation for EDs:

- An obvious recommendation is to optimize test protocols to achieve lower failure rates (e.g., better validation, fewer and/or less ambitious validity criteria) and for authorities to require tests only following protocols with a low failure rate.
- Range-finding tests can be omitted, provided that data are available from acute or chronic studies for setting appropriate concentrations in the definitive test. Although this is a reasonable assumption for fish, data are rarely available for amphibians. From acute studies, onetenth of the 96-hr LC50 can be used as the highest test concentration in the FSTRA. However, a range-finding test may be necessary if systemic toxicity or severe sublethal effects are observed near to one-tenth of the LC50 (Wheeler et al. 2013). Although omitting the range-finding test would result in

### Table 1. Estimated number of animals required for ED testing of registered pesticide and biocide active substances, based on range-finding and definitive testing, and adjusted for failure rates

| Test                        | Mechanistic assays |          | Adversity assays |          |
|-----------------------------|--------------------|----------|------------------|----------|
|                             | FSTRA** OECD TG 229 | AMA OECD | MEOGRT OECD      | LAGDA OECD |
| Number of animals in the    | 40                 | 160      | 630              | 120      |
| range-finding test          |                    |          |                  |          |
| Number of animals in the    | 96                 | 320      | 1764             | 480      |
| definitive test**           |                    |          |                  |          |
| Failure rate for definitive | 0.1                | 0.1      | 0.5              | 0.2      |
| test                        |                    |          |                  |          |
| Total number of animals in  | 128                | 512      | 3276             | 776      |
| test                        |                    |          |                  |          |
| Number of animals for 475   | 60,800             | 243,200  | 1,556,100        | 368,600  |
| pesticides                  |                    |          |                  |          |
| Total                       | 304,000            | 1,924,700| 1,924,700        | 3,034,472|
| Number of animals for 297   | 38,016             | 152,064  | 972,972          | 230,472  |
| biocides                    |                    |          |                  |          |
| Total                       | 190,080            | 1,203,444| 1,203,444        |          |

**AMA = Amphibian Metamorphosis Assay; ED = endocrine disrupting; FSTRA = Fish Short-Term Reproduction Assay; LAGDA = Larval Amphibian Growth and Development Assay; MEOGRT = Medaka Extended One-Generation Reproduction Test.**

Numbers are provided for zebrafish; using fathead minnow or medaka as test species would increase the number of animals by 30%.

**The number of animals for the definitive test was based on the minimum number of animals per replicate, the number of replicates per treatment and control, and the number of treatments (cf. the respective guidelines).**

**The failure rate describes the probability of test repetition due to inability to meet guideline validity criteria.**
a 31–36% reduction in the number of animals used in an FSTRA, this omission is rarely possible for the other tests.

- Make better use of embryo assays and mechanistic mammalian data. Embryo assays can provide in vivo mechanistic information. Fish embryo assays currently undergoing Organisation for Economic Co-operation and Development (OECD) validation can detect substances acting through the Estrogen, Androgen or Steroidogenesis (EAS) modalities (detection of Endocrine Active Substances acting through estrogen receptors, using Zebrafish embryos [EASY]; Rapid Estrogenic Activity Test In Vivo [REACTIV]; or Rapid Androgen Disruption Adverse Outcome Reporter Assay [RADAR]). For detecting thyroid activity, an amphibian embryo assay, Xenopus Eleutheroembryonic Thyroid Assay (XETA), has been adopted by the OECD. These assays are considered as “nonanimal” tests within the EU because they use embryos at a development stage where independent feeding has not yet started, whereas both the United Kingdom Animals (Scientific Procedures) Act 1986 and the Directive 2010/63/EU on the protection of animals used for scientific purposes apply to only independently feeding larval forms. The embryo assays used for investigating endocrine mechanisms therefore appear as an ethical alternative to tests conducted with older larval stages, juveniles, or adults (Halder et al. 2010). Incorporating embryo assays into testing strategies for ED testing is consistent with EU legislative animal welfare objectives, if the FSTRA or AMA were conducted only when embryo assays indicate endocrine activity.

- Avoiding duplication of mechanistic studies across vertebrates due to the high level of conservation of the endocrine system and receptor homology, as well as the key enzymes involved, extrapolation of qualitative screening level information among vertebrates is warranted. There is a high concordance between the results of amphibian or fish and rat assays for substances that interact with Estrogen, Androgen, Thyroid, Steroidogenesis (EATS) modalities (Ankley and Gray 2013; Pickford 2010). This has been confirmed by comparing protein sequence and/or structural information across species at the level of the primary amino acid sequence and functional domains (LaLone et al. 2013). Therefore, it is questionable whether additional mechanistic studies for aquatic vertebrates are needed when endocrine activity has been investigated sufficiently in mammalian assays.

Overall, the evaluation of ED properties for pesticides and biocides will result in a vast increase of the number of animals used in testing. However, utilizing cross-species extrapolation considerably reduces animal use, if no further tests on fish and amphibians are required when endocrine activity is sufficiently investigated in mammalian assays. In vivo embryo assays offer an ethical alternative to corroborate mammalian results and robustly address environmental protection goals for aquatic vertebrates.

**REFERENCES**

Ankley GT, Gray LE. 2013. Cross-species conservation of endocrine pathways: A critical analysis of tier 1 fish and rat screening assays with 12 model chemicals. Environ Toxicol Chem 32:1084–1087.

Burden N, Gellaty N, Benstead R, Benyon K, Blikk M, Cloak M, Doyle I, Edwards P, Handlej J, Katsiakadi I et al. 2017. Reducing repetition of regulatory vertebrate ecotoxicology studies. Integr Environ Assess Manag 13:955–957.

[EC] European Commission. 2016. EU Pesticides database. [accessed 2019 Jan 18]. http://ec.europa.eu/food/plant/pesticides/eu-pesticides-data/
base/public/?event=activesubstance.selection&language=EN

[ECHA] European Chemicals Agency. 2019. Information on biocides. [accessed 2019 Jan 18]. https://echa.europa.eu/de/information-on-chemi-
cals/ biocidal-active-substances

[ECHA–EFSA–JRC] European Chemicals Agency, European Food Safety Au-
thority, Joint Research Centre. 2018. Guidance for the identification of endocrine disruptors in the context of Regulations (EU) No 528/2012 and (EC) No 1107/2009.

[EU] European Union. 2017. Commission Delegated Regulation (EU). 2017/ 2100 of 4 September 2017 setting out scientific criteria for the determi-
nation of endocrine-disrupting properties pursuant to Regulation (EU) No 528/2012 of the European Parliament and Council. [accessed 2019 Jan 18]. http://data.europa.eu/eli/reg_del/2017/2100/oj

[EU] European Union. 2018. Corrigendum to Commission Regulation (EU) 2018/605 of 19 April 2018 amending Annex II to Regulation (EC) No 1107/2009 by setting out scientific criteria for the determination of endocrine disrupting properties (OJ L 101, 20.4.2018). [accessed 2019 Jan 18]. http://data.europa.eu/eli/reg/2018/605/oj

Halder M, Léonard M, Iguchi T, Ors JT, Ryder K, Belanger SE, Braunbeck TA, Embry MR, Whale G, Norberg-King T et al. 2010. Regulatory aspects on the use of fish embryos in environmental toxicology. Integr Environ Assess Manag 6:484–491.

LaLone CA, Villeneuve DL, Burgoon LD, Russom CL, Helgen HW, Berninger JP, Tierge JE, Severson MN, Cavillon JE, Ankley GT. 2013. Molecular target sequence similarity as a basis for species extrapolation to assess the ecological risk of chemicals with known modes of action. Aquat Toxicol 144–145:141–154.

[OECD] Organisation for Economic Co-operation and Development. 2009. Test No. 231: Amphibian metamorphosis assay. OECD Guidelines for the Testing of Chemicals, Section 2. Paris (FR): OECD Publishing. [accessed 2019 Jan 18]. https://doi.org/10.1787/9789264076242-en

[OECD] Organisation for Economic Co-operation and Development. 2012. Test No. 229: Fish short term reproduction assay. OECD Guidelines for the Testing of Chemicals, Section 2. Paris (FR): OECD Publishing. [accessed 2019 Jan 18]. https://doi.org/10.1787/9789264185265-en

[OECD] Organisation for Economic Co-operation and Development. 2015a. Test No. 240: Medaka extended one generation reproduction test (MEOGRT). OECD Guidelines for the Testing of Chemicals, Section 2. Paris (FR): OECD Publishing. [accessed 2019 Jan 18]. https://doi.org/10.1787/9789264242258-en

[OECD] Organisation for Economic Co-operation and Development. 2015b. Test No. 241: The larval amphibian growth and development assay (LAGDA). OECD Guidelines for the Testing of Chemicals, Section 2. Paris (FR): OECD Publishing. [accessed 2019 Jan 18]. https://doi.org/10.1787/9789264242340-en

Pickford DB. 2010. Screening chemicals for thyroid-disrupting activity: A critical comparison of mammalian and amphibian models. Crit Rev Toxicol 40:845–892.

Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market and repealing Council Directives 79/117/EEC and 91/414/EEC. OJ L 309, 24.11.2009, p 1–50.

**ORCID**

Lennart Weltje https://orcid.org/0000-0002-5191-4158
CAN STATISTICAL ANALYSES ACCOUNT FOR THE EFFECT OF INTERRELATEDNESS IN ASSESSMENTS USING MULTIPLE ENVIRONMENTAL PARAMETERS?

Keith M Somers,* † Bruce W Kilgour, † and Kelly R Munkttrick†
*Kilgour & Associates Ltd, Ottawa, Ontario, Canada
†Centre for Cold Regions & Water Science, Wilfrid Laurier University, Waterloo, Ontario, Canada
*ksomers@kilgourassociates.com
DOI: 10.1002/ieam.4166
© 2019 SETAC

Two Canadian federal monitoring programs use data on multiple indicators (or metrics) to evaluate aquatic habitat condition. The Environmental Effects Monitoring (EEM) Program requires the collection of field data on sentinel fish populations and benthic macroinvertebrates (surrogate indicator for fish habitat) to evaluate the effects of pulp and paper, metal mining, and other industrial effluents on the condition of fish populations and fish habitat (see Environment Canada 2012a). The EEM assessment uses a univariate approach to determine statistically significant differences between and among reference- and exposure-site means for 9 metrics, one at a time. A difference in any 1 metric triggers an enhanced monitoring protocol. The Canadian Aquatic Biomonitoring Network (CABIN) monitors benthic macroinvertebrates to evaluate aquatic habitat conditions (Environment Canada 2012b) but instead uses a multivariate ordination of benthic community densities and evaluates individual exposure (test) sites relative to confidence ellipses (or confidence intervals) around minimally impaired reference sites often using 3 ordination axes, 2 at a time (i.e., a bivariate approach using axes 1 and 2, axes 2 and 3, and axes 1 and 3).

Although Huebert et al. (2011) identified 3 areas for improvement in the EEM program (i.e., pseudoreplication, calculation of the Bray-Curtis Index, and the designation of α), only the α issue is discussed here. The problem arises because 9 biological metrics are evaluated individually with an α set at 0.10. This issue is known as multiplicity and it results in the increased likelihood of finding a statistically significant difference when testing multiple hypotheses (or multiple endpoints) in 1 study (Streiner 2015). In this case, the probability of a Type 1 (false-positive) error is 1 – (0.9)9 = 0.61, not 0.10. The use of a Bonferroni correction would maintain the likelihood of a false-positive error at 0.10 across all 9 tests, by changing α from 0.10 to α / n = 0.10/9 = 0.011 (Huebert et al. 2011). This proposal was criticized, as the Bonferroni correction reduces statistical power and increases the likelihood of a Type 2 (false-negative) error (Bosker et al. 2012). The Bonferroni correction is 1 of several possible options that addresses the issue of multiplicity. We will briefly discuss this and other options and recommend another solution to this issue.

The Bonferroni correction (and stepwise variants) maintains the likelihood of a Type 1 error across multiple tests (Huebert et al. 2011) but sacrifices power (Bosker et al. 2012). Another method uses the False Discovery Rate (FDR), to address the decreased power issues. However, both of these methods fail to adjust for the potential interrelatedness of the metrics (Streiner 2015). Although procedures to adjust for correlated metrics in the FDR approach have been developed, their use is not widespread. Figure 1 illustrates the consequences of failing to recognize the effect of metric interrelatedness on confidence regions, using a hypothetical data set.

The multivariate analysis of variance (MANOVA) provides a solution to problems with multiple tests and correlated metrics (Lehmacher et al. 1991). In MANOVA, all metrics are evaluated simultaneously, resolving the multiple-comparison, false-positive issue using a single test. Correlations among metrics are also incorporated into MANOVA, so that the issue of doublecounting redundant metrics is addressed. However, because MANOVA evaluates all metrics simultaneously, a different null hypothesis (Ho) is tested (i.e., Ho: all metric differences are 0). As a result, rejection of a multivariate null hypothesis may require secondary tests (see Lehmacher et al. 1991) and additional statistical tools (e.g., discriminant functions or canonical variates analysis) to interpret the result and identify important impacted metrics (e.g., see Bowman and Somers 2006). These secondary tests could be used to evaluate alternate hypotheses based on specific response patterns, inferring the cause of the impairment.

For regulatory programs like EEM and CABIN, evaluation of test sites using thresholds based on the normal range of reference-site variation is possible using univariate and multivariate approaches (Bowman and Somers 2006). In the univariate case, the difference between the reference-area mean and exposure-area mean (or a reference-area mean and a test site) is compared to a threshold based on the standard deviation, but without compensation for the effect of correlations between metrics (i.e., the confidence region for 2 metrics considered separately can differ appreciably from a confidence region that accounts for the correlation between 2 metrics; see Figure 1). In the multivariate scenario, the difference between reference- and exposure-area
means (i.e., centroids, or the reference-area centroid and a test site) is compared to a threshold that is a function of all of the metrics, their correlations, and the multivariate reference-area variation to produce an \( F \)-value and associated probability much like the univariate case. A significant MANOVA can be followed with discriminant functions analysis and partial analyses (see Bowman and Somers 2006), secondary MANOVAs testing specific alternate hypotheses, or ANOVAs with a Bonferroni correction (Huebert et al. 2011).

In summary, the use of traditional MANOVA-based approaches by EEM and CABIN should provide more rigorous data analyses that result in identification of true environmental impairments that signal a need for enhanced monitoring. This statistical approach is valid and useful for any area of research that includes a multiplicity of metrics.

REFERENCES
Bosker T, Barrett TJ, Munkittrick KR. 2012. Response to Huebert et al. (2011), Canada’s environmental effects monitoring program: areas for improvement. Integr Environ Assess Manag 8:381–382.
Bowman MF, Somers KM. 2006. Evaluating a novel Test Site Analysis (TSA) bioassessment approach. J N Am Benthol Soc 25:712–717.

Figure 1. Four different data sets (A–D) are presented. These data sets consist of 30 reference-site data points (yellow-filled circles), drawn from multivariate normal populations with increasing specified correlations (i.e., Pearson’s \( r \) = A: 0.0, B: 0.3, C: 0.6, and D: 0.9) between two standardized metrics (i.e., means of 0 and SD of 1). The area within the dashed square represents the 95% confidence region defined by the normal range (i.e., mean \( \pm t \times SD \); Bowman and Somers 2006) when both metrics are independent, whereas the ellipse (solid line) is the 95% confidence region for both metrics, considered simultaneously. When the correlation between metrics is ignored, the test site (red-filled circle) lies within the normal range for each metric and it is a constant Euclidean distance (ED = 1.41) from the reference-site mean (black-filled circle). By contrast, the ellipse and generalized distance (GD) incorporate the correlation between metrics and change as the correlation changes. The univariate approach uses the dashed square and ED, whereas the multivariate approach uses the ellipse and GD to determine whether a test site falls outside of the normal range (after Lehmacher et al. 1991).
Environment Canada. 2012a. Environmental effects monitoring technical guidance. Gatineau (QC). 550 p.
Environment Canada. 2012b. Canadian aquatic biomonitoring network field manual for wadeable streams. Dartmouth (NS). 57 p.
Huebert DB, McKernan M, Paquette C. 2011. Canada’s environmental effects monitoring program: Areas for improvement. Integr Environ Assess Manag 7:143–144.
Lehmacher W, Wassmer G, Reitmeir P. 1991. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. Biometrics 47:511–521.
Streiner DL. 2015. Best (but oft-forgotten) practices: The multiple problems of multiplicity – Whether and how to correct for many statistical tests. Am J Clin Nutr 102:721–728.