2016-11

Optimizing the design of a reproduction toxicity test with the pond snail Lymnaea stagnalis.

Charles, S

Regulatory toxicology and pharmacology : RTP

This is an accepted manuscript of an article published by Elsevier in Regulatory Toxicology & Pharmacology, 25 July 2016. Available at DOI: 10.1016/j.yrtph.2016.07.012

Running head:

Optimizing reproduction toxicity test for aquatic molluscs

Corresponding author:

Professor Sandrine CHARLES

Laboratoire de Biométrie - Biologie Evolutive

Université de Lyon; Université Lyon 1;

CNRS; UMR 5558;

Bâtiment Gregor Mendel, Mezzanine

43 boulevard du 11 novembre 1918

F-69622 Villeurbanne Cedex, France

Tel. +33 (0)4 7243 2900

Mail: sandrine.charles@univ-lyon1.fr

Optimizing the design of a reproduction toxicity test with the pond snail *Lymnaea stagnalis*

Authors

CHARLES Sandrine[†], DUCROT Virginie[‡,§], AZAM Didier[‡], BENSTEAD Rachel[||], BRETTSCHNEIDER Denise[#], DE SCHAMPHELAERE Karel[††], FILIPE GONCALVES Sandra [‡‡], GREEN John W.[§§], HOLBECH Henrik[|| ||], HUTCHINSON Thomas H.[##], FABER Daniel[§], LARANJEIRO Filipe[†††], MATTHIESSEN Peter[‡‡‡], NORRGREN Leif[§§§], OEHLMANN Jörg[#], REATEGUI-ZIRENA Evelyn[|| || ||], SEELAND-FREMER Anne[###], TEIGELER Matthias[††††], THOME Jean-Pierre[‡‡‡‡], TOBOR KAPLON Marysia[§§§§], WELTJE Lennart[|| || || ||], LAGADIC Laurent[‡,§]

Affiliations

† Univ Lyon, Université Lyon 1, UMR CNRS 5558, Laboratoire de Biométrie et Biologie Évolutive, F-69100 Villeurbanne, France

‡ Institut National de la Recherche Agronomique (INRA), Centre de Recherche de Rennes, 65 rue de Saint-Brieuc, F-35042 Rennes, France

§ Bayer Aktiengesellschaft, Crop Science Division, BCS AG-R&D-D-EnSa-ETX-AQ, Alfred-Nobel Straße 50, D-40789 Monheim am Rhein, Germany

|| The Food and Environment Research Agency (now Fera Science Ltd). Sand Hutton, York, YO41 1LZ, United Kingdom.

Goethe University Frankfurt am Main, Department Aquatic Ecotoxicology, Max-von-Laue-Straße 13, D-60438 Frankfurt, Germany

†† Laboratory of Environmental Toxicology and Aquatic Ecology, Faculty of Bioscience Engineering, Ghent University, Belgium

‡‡ Department of Biology & CESAM, Centre for Environmental and Marine Studies, University of Aveiro, 3810-193 Aveiro, Portugal

§§ DuPont, PO Box 60, 1090 Elkton Road, DuPont Stine-Haskell Research Center, S315/1369, Newark, Delaware, USA.

|| || Department of Biology, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark

School of Biological Sciences, University of Plymouth, Plymouth PL4 8AA United Kingdom

††† Departamento de Biologia, Universidade de Aveiro, 3810-193 Aveiro, Portugal

‡‡‡ Old School House, Brow Edge, Backbarrow, Ulverston, Cumbria LA128QX, United Kingdom

§§§ Department of Pathology Faculty of Veterinary Science Swedish University of Agricultural Sciences P.O. Box 7028 Uppsala, S-750 07 Sweden

|| || || Department of Environmental Toxicology, Texas Tech University, Lubbock, TX, USA.

Ibacon GmbH; Arheilger Weg 17, 64380 Rossdorf, Germany

†††† Fraunhofer Institute for Molecular Biology and Applied Ecology, Department of Ecotoxicology, Auf dem Aberg 1, 57392 Schmallenberg, Germany.

‡‡‡‡ University of Liège, Laboratory of Animal Ecology and Ecotoxicity (LEAE-CART), Allée du 6 Août, 11, Sart-Tilman, Belgium

§§§§ WIL Research, Department of In vitro and Environmental Toxicology, Ashland, United States

|| || || || BASF SE, Crop Protection – Ecotoxicology, Speyerer Straße 2, D-67117 Limburgerhof, Germany

Abstract

This paper presents the results from two ring-tests addressing the feasibility, robustness and reproducibility of a reproduction toxicity test with the freshwater gastropod *Lymnaea stagnalis* (RENILYS strain). Sixteen laboratories (from inexperienced to expert laboratories in mollusc testing) from nine countries participated in these ring-tests. Survival and reproduction were evaluated in *L. stagnalis* exposed to cadmium, tributyltin, prochloraz and trenbolone according to a draft OECD Test Guideline. In total, 49 datasets were analysed to assess the practicability of the proposed experimental protocol, and to estimate the between-laboratory reproducibility of toxicity endpoint values. The statistical analysis of count data (number of clutches or eggs per individual-day) leading to ECx estimation was specifically developed and automated through a free web-interface, allowing users to reproduce the whole analysis. Based on a complementary statistical analysis, the optimal test duration was established and the most sensitive and cost-effective reproduction toxicity endpoint was identified, to be used as the core endpoint. This validation process and the resulting optimized protocol were used to consolidate the OECD Test Guideline for the evaluation of reproductive effects of chemicals in *L. stagnalis*.

Keywords

Mollusc, Fecundity, ECx, Count data, Test design optimization

Introduction

In 2010, the Organization for Economic Cooperation and Development (OECD) recommended the development of a new test guideline for reprotoxicity testing in freshwater molluscs [1]. Between 2011 and 2013, a 56-days reproductive semi-static-renewal test protocol was evaluated in a prevalidation ring-test using *Lymnaea stagnalis* (Linnaeus, 1758) and involved seven laboratories in Europe [2]. Subsequent statistical analyses provided robust estimates of x% lethal and effective concentrations (LCx and ECx) for both clutch- and egg-based endpoints, and between-laboratory comparison demonstrated a low variability in LCx and ECx values. In addition, a consolidated draft of the standard operating protocol was provided with detailed rearing and toxicity test procedures as well as their application to evaluate reproductive toxicants [2]. Consequently, both the OECD Validation Management Group for Ecotoxicity testing (VMG-Eco) and the OECD ad-hoc Expert Group on Invertebrate Testing further supported a validation ring-test.

The aim of the validation ring-test was threefold: (i) assessing the reproducibility of the test results among a larger number of laboratories with different levels of experience in mollusc testing (from inexperienced to experts); (ii) assessing consistency and reproducibility of toxicity thresholds (i.e., ECx values estimated for all laboratories) between the two ring-tests (i.e., prevalidation vs. validation steps); (iii) assessing responses of snails to a larger number of chemicals. In addition, key issues related to optimization of the test design also deserved elucidation: (i) costs, benefits and feasibility in reducing the exposure duration (i.e., could the test duration be reduced while safeguarding accuracy and precision of ECx estimates?); (ii) benefits of recording both the number of clutches and the number of eggs per clutch (i.e., does the choice of the recorded endpoint matter when estimating toxicity thresholds?).

The validation ring-test was conducted from October 2013 to October 2014 according to the draft standard operating procedure. In total, 13 laboratories from academia, government, industry and consultancy, in Europe and North-America, participated in collecting raw data and water samples for statistical and chemical analyses, respectively. Six laboratories (all new compared to the laboratories involved in the prevalidation ring-test) were in charge of testing cadmium (Cd), which had been used in the prevalidation ring-test [2]. Five laboratories tested tributyltin (TBT), four laboratories tested prochloraz (PRO), and two laboratories tested trenbolone (TRB). The choice of these substances was based upon recommendations from the OECD VMG-Eco (Table 1). They were assumed to cause adverse effects on snail reproduction (as confirmed in pre-tests that were conducted for all chemicals except trenbolone). These substances reflect different levels of complexity in terms of toxicity testing; Cd is an "easy-to-test" substance, whereas TBT, PRO and TRB are more difficult substances to test (e.g., use of solvent required for TBT; limited stability of PRO in water, both difficulties being encountered for TRB [3]). Hence, performing the validation ring-test with difficult test substances could contribute to further demonstrate the robustness of the experimental protocol and to identify the most relevant reproduction endpoint in *L. stagnalis*.

This paper presents the results of the validation ring-tests for Cd, TBT and PRO, in comparison with those of the prevalidation ring test where applicable. Exposure of snails to TRB up to a mean measured concentration of 776 $ng.L^{-1}$ had no effects on their reproduction; the corresponding results are thus not presented in this paper. For the remaining substances, ECx were estimated for each laboratory and then compared in order to assess their reproducibility between laboratories. We also investigated the consequence of reducing the exposure duration on both ECx median value and uncertainty. Finally, after having confirmed the low between-laboratory variability

when reducing the exposure duration, we considered the possibility of recording only one core endpoint to be used in the OECD test guideline for the reproduction toxicity tests with *L. stagnalis*.

Materials and Methods

Implementation of the validation ring-test

The experimental design used to collect raw data during the validation ring-test followed the one used for the prevalidation ring-test. All details about test organisms, snail acclimation, tested chemicals, experimental conditions, sampling and analysis of test media, and collection of raw data are available in Ducrot et al. [2] and summarized in Supplementary Information (Table S0). The principle of the reproduction toxicity test and the specificities of the validation ring-test are here recalled.

Principle of the reproduction toxicity test

The primary objective of the test was to assess the effect of chemicals on the reproductive output of *L. stagnalis*. To this end, reproducing adults of *L. stagnalis* were exposed to a range of 5 concentrations of the test chemical and a control (water only or, when required, a solvent control) and monitored for 56 days for survival and reproduction. No less than 6 replicates of 5 snails were exposed to each concentration (i.e., 30 snails per treatment and per control). Prior to the test, snails were sampled from a laboratory parasite-free culture, checked for identical size (27 ± 2 mm), and introduced into test vessels for a few days acclimation period. As soon as exposure to the test chemical started (i.e., day 0 of the test), survival and fecundity were recorded at least twice a week, before feeding the snails *ad libitum* with (organic) round-headed lettuce and renewing water. Dead snails were counted and withdrawn from the test vessels. Both the number of clutches and the number of eggs per clutch were counted.

Raw data were collected in a spreadsheet automatically providing a text file under the appropriate format for the statistical analyses.

Tested chemicals and exposure water sampling and analysis

Specifications of the test chemicals are provided in Table 1. Nominal concentrations for Cd were chosen based on the prevalidation ring-test, namely 25, 50, 100, 200, 400 $\mu g.L^{-1}$. Nominal concentrations were 87.5, 175, 350, 700, 1400 $ng.L^{-1}$ and 10, 32, 100, 320, 1000 $\mu g.L^{-1}$ for TBT and PRO, respectively. Water samples were collected before and after water renewal, at the beginning, mid-term and end of each experiment for the determination of actual exposure concentrations (42 samples per experiment). Actual Cd concentrations in water were measured in 50 mL acidified samples (triplicates) by atomic adsorption spectrometry (limit of detection: 0.8 $\mu g.L^{-1}$). Actual TBT concentrations in water were measured in triplicate by coupled capillary gas chromatography to mass spectrometry (GC-MS-MS; ITQ100, Thermo Scientific, USA) according to Giusti et al. [4] with slight modifications. The limit of detection (LOD) was 6 ng $TBT.L^{-1}$ and the limit of quantification (LOQ) was 18 ng $TBT.L^{-1}$ (concentrations are expressed in ng $TBT.L^{-1}$: equivalent in ng $Sn.L^{-1}$ can be calculated by dividing these values by a factor 2.44). The mean recovery efficiency was 99% ± 18.6% and was in good agreement with requirements of the SANCO guidance document [5]. PRO samples were analysed directly from filtered samples by LC-MS-MS (LOD: 3.9 $\mu g.L^{-1}$, LOQ: 1.56 $\mu g.L^{-1}$ and mean recovery efficiency: 70%± 6.3%).

Statistical modelling of reproduction data

Solvent controls were used as the reference for statistical analysis of the TBT data (all laboratories). We used the Jonckheere-Terpstra hypothesis test as a way to discriminate datasets for which the chemicals had a significant effect on the reproduction endpoints.

186 The Jonckheere-Terpstra hypothesis test was here performed under the R software [6]
187 with package 'clinfun' and function 'jonckheere.test' [7]; alternatively, the FREQ SAS
188 procedure or other software may have been used. With R package 'clinfun', it was not
189 possible to run an exact Jonckheere-Terpstra hypothesis test due to ties in some
190 datasets. We thus used the normal approximation with a fixed number of 10^6 iterations.
191 Statistical modelling of reproduction data was performed in order to estimate ECx
192 values. ECx estimation was performed under the R software [6] with package 'morse'
193 [8], according to the new approach proposed by Delignette-Muller et al. [9], both taking
194 into account mortality among parents without losing valuable data and describing
195 potential between-replicate variability. All the statistical analyses presented in this
196 paper are identically reproducible using the free web-platform MOSAIC and its module
197 MOSAIC_repro [10]. Raw data were analysed using the same procedure for both
198 reproduction endpoints (number of clutches or number of eggs per clutch), as explained
199 below.
200 *Calculation principle of the number of individual-day*
201 A non-negligible mortality may be recorded in exposed snails at the end of the test, due
202 to the prolonged exposure duration (56 days) chosen to investigate optimal test
203 duration. Nevertheless, individuals may have reproduced before dying and thus have
204 contributed to the cumulative reproduction outcome observed at the end of the test.
205 Information on the reproduction of individuals which died during the test, should
206 therefore be taken into account to avoid any bias in the statistical analyses. This is
207 particularly critical at high exposure concentrations, where mortality may be high.
208 In the *L. stagnalis* reproduction toxicity test, mortality was regularly recorded at each
209 time-point when clutches (resp. eggs) were counted. The period during which each
210 individual was alive, corresponding to the period during which it may have reproduced,

211 could thus be determined. As commonly done in epidemiology for incidence rate
212 calculations, it was possible to calculate, for one replicate, the sum of the observation
213 periods of each individual before its death. When an organism was alive at time t but
214 counted as dead at time (t + 1), it was then assumed to be actually dead at ((t + 1) + t)/2.
215 The final sum for a replicate can then be expressed as a number of individual-days for
216 the respective replicate. Hence, reproduction was expressed for each replicate as the
217 number of clutches (resp. the number of eggs per clutch) per individual-day.
218 *Fit principle of the regression model*
219 Let N_{ij} be the number of offspring (clutches or eggs per clutch) for replicate j at the i^{th}
220 concentration u_i, and NID_{ij} the number of individual-days at the i^{th} concentration for
221 replicate j. As a first approximation, if the possible between-replicate variability is
222 neglected, a Poisson distribution can describe N_{ij}:

223 $$N_{ij} = Poisson\left(f\left(u_i;q\right)\times NID_{ij}\right) \qquad (1)$$

224 where $f\left(u_i;q\right)$ is the deterministic part of the model describing the mean tendency of
225 the exposure-effect relationship.
226 Depending on the dataset, several deterministic parts may be suitable: the 3, 4 or 5-
227 parameter log-logistic models, the Gompertz model, the 2 or 3-parameter exponential
228 models, the Bruce-Versteeg model or the Brain-Cousens model [11]. In this paper, for
229 our comparison needs between laboratories, we chose the three-parameter log-logistic
230 model which appeared as describing at best the mean tendency in most of the datasets:

231 $$f\left(u_i;q\right) = \frac{d}{1+\left(u/EC_{50}\right)^b} \qquad (2)$$

232 where $q=\left(EC_{50},d,b\right)$, EC_{50} is the concentration inducing a halfway effect between upper
233 limit d and 0, while b stands for the shape of the curve.

In order to explicitly account for the between-replicate variability, the previous Poisson model may be extended with a gamma distribution [9]:

$$N_{ij} \sim Poisson\left(f_{ij} \times NID_{ij}\right) \quad \text{with} \quad f_{ij} \sim gamma\left(\frac{f\left(u_i;q\right)}{w};\frac{1}{w}\right) \qquad (3)$$

where parameters $f\left(u_i;q\right)$ and $wf\left(u_i;q\right)$ are respectively the mean and the variance of the gamma distribution. Parameter w corresponds to an over-dispersion parameter (the greater its value, the greater the between-replicate variability).

Because non-standard stochastic parts (Poisson or gamma-Poisson) were required, we chose the Bayesian framework to infer parameter estimates from experimental data. For that purpose, we chose the R package 'morse' [8] that proposes the combined use of freeware JAGS [12] and software R [6]; alternatively SAS MCMC procedures or the WinBUGS software may also be used. Both models (Poisson and gamma-Poisson) were systematically fitted on each dataset, and the Deviance Information Criterion (DIC) was used to choose the most appropriate stochastic part of the model. In situations where over-dispersion (that is between-replicate variability) could be neglected, the Poisson model provided more reliable estimates (with narrower credible intervals). Hence a Poisson model was preferred unless the gamma-Poisson model had a significantly lower DIC (in practice we required a difference of 10).

The use of Bayesian inference requires the choice of appropriate priors based on expert knowledge on *L. stagnalis* reproduction process and the experimental design itself:

- $\log_{10}\left(EC_{50}\right) \sim N\left(m;s\right)$ where m and s are defined from $u_{\min}$ and $u_{\max}$, that is the minimum (excluding the control) and the maximum tested concentrations, respectively, as follows:

256 $$m = \frac{\log_{10}\left(u_{\min}\right) + \log_{10}\left(u_{\max}\right)}{2} \text{ and } s = \frac{\log_{10}\left(u_{\max}\right) - \log_{10}\left(u_{\min}\right)}{4}$$

257 We thus assumed a normal distribution for $\log_{10}\left(EC_{50}\right)$ centred on the mean of

258 $\log_{10}\left(u_{\min}\right)$ and $\log_{10}\left(u_{\max}\right)$, with the probability that $\log_{10}\left(EC_{50}\right)$ lies between

259 $\log_{10}\left(u_{\min}\right)$ and $\log_{10}\left(u_{\max}\right)$equals to 0.95;

260 • As *d* stands for the reproduction output in controls, we set a normal prior

261 $N\left(m_d, s_d\right)$ based on the data themselves:

262 $$m_d = \frac{1}{r_0}\sum_j \frac{N_{0j}}{NID_{0j}} \text{ and } s_d = \sqrt{\frac{\sum_j \left(\frac{N_{0j}}{NID_{0j}} - m_d\right)^2}{r_0\left(r_0 - 1\right)}}$$

263 where r_0 is the number of replicates in the controls. Note that since the replicates

264 in the controls were used to define the prior distribution of *d*, they were excluded

265 from the fitting process;

266 • $\log_{10}\left(b\right) \sim U\left(-2,2\right)$ a quasi-non-informative prior for the shape parameter;

267 • $\log_{10}\left(w\right) \sim U\left(-4,4\right)$, a quasi-non-informative prior for the over-dispersion

268 parameter of the gamma-Poisson distribution.

269 The major advantage of Bayesian inference lies in the posterior distributions it provides

270 as estimates of each parameter. From there, a posterior distribution can also be

271 obtained for any ECx whatever x. Posterior distributions are usually summarised as a

272 median value and its associated 95% credible interval extracted from 2.5, 50 and 97.5%

273 quantiles, respectively. An alternative analysis was conducted based on standard models

274 of adjusted reproduction data, defined as Nreproadj = Nreprocumul/Nindtime

275 computed on a replicate basis (results not shown); this alternative analysis provided

ECx estimates very similar to those from the (gamma-)Poisson models, including those with alternative deterministic forms for the mean tendancy (results not shown).

Datasets

A full statistical analysis was conducted on all available datasets, i.e., datasets from the prevalidation ring-test [2] and datasets from the validation ring-test presented hereafter. Combining ring-tests (prevalidation and validation), endpoints (number of clutches and number of eggs per clutch) and chemicals (Cd, TBT and PRO) from the participating laboratories resulted in a total of 84 datasets to analyse. For each dataset, EC_{50} values were estimated for cumulative reproduction per individual-day over 56 days, expressed via either the number of clutches, or the number of eggs per clutch.

Optimizing the exposure duration

For each of the considered endpoints, the possible reduction in the experiment duration was investigated by comparing the EC_{50} estimates (median and 95% credible interval) obtained in a given laboratory at time 21, 28, 35, 42 and 49 days with the median EC_{50} value obtained after 56 days (denoted by $EC_{50\text{-}56d}$ hereafter) surrounded with the variability between all laboratories. This inter-laboratory variability was calculated as plus or minus the standard deviation (*sd*) of all median $EC_{50\text{-}56d}$ values, separately from the pre-validation and validation ring-tests. We considered as optimal the shortest exposure duration that was outside the inter-laboratory variability range. For this shortest exposure duration, the $EC_{50\text{-}d}$ estimate for a given laboratory at day *d* was considered as not different from the $EC_{50\text{-}56d}$ estimate, meaning that a stable enough EC_{50} estimate had been reached at day *d* already.

Analyses of the datasets from the prevalidation and validation ring-tests were handled separately because the experimental design slightly changed between the; indeed, the

validation ring-test was performed based on a consolidated draft of the standard operating protocol two (see SI, Table S0). Consequently, we used 4 different *sd* values: 2 different *sd* values for the clutch- and egg-based endpoints within the prevalidation ring-test and 2 different *sd* values for the clutch- and egg-based endpoints within the validation ring-test.

Comparing results from clutch- and egg-based endpoints

For the chosen optimal exposure duration, we investigated whether the EC_{50} could be accurately estimated based upon the number of clutches alone or whether eggs must be also counted. For that purpose, we compared the posterior probability distributions of EC_{50} values, as provided by the Bayesian inference method, using clutch and egg data. We used the R package 'fitdistrplus' [13] in order to obtain the Cullen and Frey graph. This skewness-kurtosis plot helps to choose the most appropriate distribution among common ones. Given that priors on EC_{50} were lognormally distributed, we may expect also a lognormal distribution for the posteriors. Once the suitability of the lognormal law for the posteriors was established, we used the following indices to check similarities between posterior distributions of EC_{50} estimates from clutch and egg data:

- the 2.5 and 97.5% quantiles (denoted $Q2.5_{EC50}$ and $Q97.5_{EC50}$, respectively) from the EC_{50} posterior distribution;
- the mean, standard deviation and coefficient of variation from the fitted lognormal distribution;
- the uncertainty of EC_{50} estimates, namely $Q_{extent} = Q2.5_{EC50} - Q97.5_{EC50}$.

Results

Validation ring-test results at day 56

Test validity

Test validity criteria as stated in the consolidated standard operating procedure were achieved in all laboratories: temperature remained within the 20 ± 1 °C range; oxygen saturation did not drop below 60% air saturation value (ASV; 5.4 $mg.L^{-1}$ at 20°C); mortality did not exceed 20% in control groups by the end of the test; fecundity in the controls was at least 8 egg-clutches per snail at the end of the 56d test. In addition, each laboratory was able to maintain an appropriate water quality: pH was in the 7.0 - 8.5 range; conductivity in the 400 – 800 $\mu S.cm^{-1}$ range; and water hardness was in the 140 – 250 $mg.L^{-1}$ range.

Measured exposure concentrations

Mean measured concentrations were calculated for each chemical and laboratory as the arithmetic mean of all measured values over the test duration. They were linearly related to the nominal concentration (see SI, Figure S0). Mean measured Cd concentrations (calculated for all participating laboratories) were 19, 35, 70, 149 and 300 $\mu g.L^{-1}$, which compare to nominal values of 25, 50, 100, 200 and 400 $\mu g.L^{-1}$. Mean measured TBT concentrations were 39, 78, 118, 251 and 435 $ng.L^{-1}$, which compare to nominal values of 87.5, 175, 350, 700 and 1,400 $ng.L^{-1}$. Mean measured PRO concentrations were 13, 21, 56, 324 and 765 $\mu g.L^{-1}$, which compare to nominal values of 10, 32, 100, 320 and 1,000 $\mu g.L^{-1}$. The mean measured exposure concentration values specific to each laboratory were used for the estimation of ECx values.

Test results

For all laboratories (with two exceptions) and all tested chemicals, both clutch- and egg-based endpoints significantly decreased with increasing concentrations (Jonckheere-Terpstra p-values < 0.05, Table S1). $EC_{x\text{-}56d}$ estimates (x = 10, 50) are detailed in SI (Tables S2, S3 and S4) and summarized in Figure 1 for Cd (in SI, Figures S1 and S2, for TBT and PRO, respectively).

Reproducibility of results between laboratories

The coefficients of variation of $EC_{50\text{-}56d}$ values between laboratories are given in Table 2 for all tested chemicals. They were in the range 28.0 - 52.5% for the validation ring-test, that is similar values to those obtained during the prevalidation ring-test (21.8 - 42.0%).

Optimizing the experimental design

For Cd, median $EC_{50\text{-}56d} \pm sd$ intervals used to compare EC_{50} estimates (median and 95% credible interval) at each exposure duration (from 21 to 56 days) are given in Figures 2 and 3 for the prevalidation and validation ring-tests, respectively. Results for TBT and PRO are given in Supplementary Information (Figures S3-S5).

The between-laboratory variability was less important in the prevalidation ring-test than in the validation ring-test, due to the higher expertise of participating laboratories in the prevalidation phase. This resulted in smaller median $EC_{50\text{-}56d} \pm sd$ intervals (i.e., thinner grey band). Therefore, optimal exposure duration was greater in the prevalidation ring-test (i.e., 35 days) than in the validation ring-test (i.e., 28 days) (Figures 2 and 3). Considering that the experimental protocol was in its final version for the validation ring-test (see SI, Table S0, for differences between the pre-validation and validation tests), we referred to the corresponding results to decide whether an exposure duration of 28 days would be sufficient to ensure adequate test sensitivity

Test results at day 28

All datasets corresponding to both the prevalidation and validation ring-tests were analysed simultaneously at day 28. As shown in Table S1, both endpoints were significantly altered within the tested concentration range for all laboratories whatever the chemical, except for Lab. 11 with Cd and the clutch-based endpoint (Jonckheere-Terpstra test, p-value = 0.62) and for Lab. 07 with PRO and the clutch-based endpoint

373 (Jonckheere-Terpstra test, p-value = 0.080). $EC_{x\text{-}28d}$ estimates are detailed in SI
374 (Tables S2, S3 and S4). Results show robust $EC_{50\text{-}28d}$ estimates with small uncertainty
375 and a good agreement between values obtained in the different laboratories (Table 2).
376 In addition, for all datasets, several goodness-of-fit criteria were checked, in particular
377 the comparison of prior-posterior probability distributions as well as the so-called
378 posterior predictive checks, that is plots of the observed values against their
379 corresponding estimated predictions, along with their 95% credible interval (results not
380 shown).
381 For Cd, there was less variability in the EC_{50} values estimated at day 28 than at day 56,
382 as shown by smaller coefficients of variation between laboratories at day 28. For TBT,
383 the variability was also reduced between results at day 28 and results at 56, but only for
384 the prevalidation ring-test; the high coefficient of variation values for the clutch (57.3%)
385 and the egg (63.4%) endpoints of the validation ring-test at day 28 were due to high
386 estimates of $EC_{50\text{-}28d}$ for Lab. 02 compared to those obtained at day 56 (see SI, Figure S1).
387 For TBT, low EC_{50} estimates for Lab. 03 probably also biased calculations of the
388 coefficients of variation (see SI, Figure S1). At last, for PRO, coefficients of variation were
389 similar between results at day 28 and results at day 56, as well as between both
390 endpoints.
391 To confirm that EC_{50} estimated at days 28 and 56 were close, we also calculated ratios
392 between EC_{50} medians at 28 and 56 days, as well as ratios between EC_{50} medians from
393 clutches at 28 days and EC_{50} medians from eggs at 56 days. Only three of these ratios
394 were slightly over 2 (twice for Cd, once for TBT).

Choosing the main core endpoint

Overlapping boxplots on Figure 1 (resp. Figures S1 and S2) illustrate the similarity between EC_{50} estimates from clutch and egg-based endpoints at day 28. Figure S6 strengthens this result based on the comparison of full posterior distributions of EC_{50} estimates superimposed to prior ones: distributions have similar positive skewness and similar kurtosis; peaks of distributions are also closely located in most cases.

From Table 3, we notice that EC_{50} medians and uncertainty extents (given by Q_{extend}) are very good proxies of mean and standard deviation of the fitted lognormal distribution: $\mu_{EC_{50}} \simeq \text{Median}_{EC_{50}}$ and $\sigma_{EC_{50}} \simeq Q_{extent}/4$. The coefficients of variation confirm these results with equal values from clutch- or egg-based endpoint, except in three cases out of 22 comparisons (bold numbers in Table 3). EC_{50} medians from clutches were generally similar to EC_{50} medians from eggs (Table 3); indeed, both EC_{50} medians remained similar based on EC_{50} median ratios close to 1 (except for Lab.13 with Cd in the validation ring-test).

Discussion

The feasibility, robustness and reproducibility of the protocol proposed for an OECD reproduction toxicity test guideline with *L. stagnalis* was addressed in two validation exercises (see Ducrot et al. [2] for the prevalidation ring-test and the present paper for the validation ring-test) with four different chemicals. In total, 16 laboratories (from inexperienced to expert laboratories in mollusc testing) from nine countries participated in these ring-tests.

Within these validation exercises, 23 reproduction toxicity tests were performed, among which only a few did not achieve the given validity criteria. Two laboratories had technical issues to satisfy the temperature criterion of 20°C and another laboratory had

issues in maintaining the appropriate concentration of dissolved oxygen in test water. Such technical issues could easily be fixed. In addition, these three laboratories did not meet the biological criteria (maximum control mortality or minimum clutch number in control groups) as established during the prevalidation ring-test. Minimum clutch number in control groups was set to the lowest value obtained in the prevalidation ring-test to ensure that the presently given test validity criteria are appropriate and achievable.

For all tested chemicals, results of the reproduction tests were estimated with good precision, i.e., small 95% credible intervals, indicating that the test protocol and method used to estimate the EC_x values were robust. Results were also homogenous between laboratories, since most of the laboratories provided comparable EC_{10} (see Table S2-S4) and EC_{50} values with overlapping 95% credible intervals (Figure 1). For Cd and TBT (with the exception of Lab. 08), a 2-fold difference was obtained between the lowest and the highest estimated $EC_{50\text{-}56d}$ values (using either the number of clutches, or the number of eggs per individual-day). For Cd, lower $EC_{50\text{-}56d}$ values were found for both endpoints in Lab. 02. The softness of test water used in this laboratory (< 50 $mg.L^{-1}$ of $CaCo_3$) may explain this result, as water softness is known to increase the Cd toxicity [14]. A similar trend was already observed in the prevalidation ring-test (see Lab. 07 Figure 1, [2]). The high coefficient of variation value for the clutch-based endpoint with Cd in the validation ring-test (52.5%) was due to the high estimate of $EC_{50\text{-}56d}$ for Lab. 11 (Figure 1). For PRO, inter-laboratory variability in EC_{50} values was below a factor 2. These results attest to a good reproducibility of the $EC_{50\text{-}56d}$ values between laboratories. Indeed, these differences are in the range of acceptable variation defined for reference chemicals in OECD guidelines for acute toxicity tests with invertebrates (i.e., factor 3.5 for $K_2Cr_2O_7$ in TG 202 and factors 3.5 and 7.2 in TG 235 for KCl and 3,5-DCP, respectively

444 [11, 12]). Obtaining consistent endpoint values among all laboratories and when
445 repeating the ring-tests demonstrates the robustness of the proposed test protocol, as
446 well as the reproducibility of derived results.
447 EC_{50} values estimated based on the number of clutches per individual-day did not
448 significantly differ from EC_{50} values estimated based on the number of eggs per
449 individual-day: both endpoints were equally sensitive for all tested chemicals.
450 Therefore, both endpoints could be used in the reproduction toxicity tests with *L.*
451 *stagnalis*. However, assessing only the number of clutches produced per individual-day
452 is sufficient to obtain robust EC_{50} estimates. Indeed, the ratio between median EC_{50}
453 values estimated based on clutches vs. eggs is close to 1 for all laboratories, except
454 Lab. 13 where it reached a value of 2 (Table 3).
455 EC_{50} values estimated based on either clutches or eggs per individual-day after 28 and
456 56 days did not significantly differ, for any of the tested chemicals. Indeed, the mean
457 ratio (for all laboratories, endpoints, chemicals, and the two ring-tests) between median
458 EC_{50} values estimated at 28 days vs. 56 days was 1.2 (Table S5). The highest difference
459 was found in Lab. 02 where it reached a value of 2.1 during the TBT validation test and
460 using the clutch-based endpoint (Figure 1). For Cd and TBT, inter-laboratory variability
461 in EC_{50} values was smaller at 28 days compared to 56 days, as shown by smaller values
462 of the between-laboratory coefficient of variation at 28 days, while the same between-
463 laboratory variability was observed after 28 days vs. 56 days for PRO. Based on these
464 results, the test duration could be reduced to 28 days without hampering the accuracy of
465 the EC_{50} estimate.
466 Overall, the above-mentioned results suggest that the test duration can be reduced from
467 56 days to 28 days, and the number of clutches per individual-day can be used as the
468 core measure for the reproductive output (instead of counting all eggs) with no

469 influence on the accuracy and precision of EC_{50} estimate. To further strengthen this
470 assumption, we calculated the ratio between EC_{50} values obtained under the optimized
471 test design (28 d, using clutch number as a measure for the reproductive output) and
472 those obtained using the non-optimized test design (56 d, using egg number as a
473 measure of the reproductive output). This calculation was performed for all laboratories
474 and chemicals and for both the validation and prevalidation ring-tests. The obtained
475 mean ratio was 1.3 showing that, on average, the median EC_{50} estimate obtained with
476 the optimized design was 1.3 fold lower than the median EC_{50} estimate obtained with
477 the non-optimized design. The maximal difference was estimated to be a ratio of 2.7
478 (obtained in Lab. 13 for the Cd validation test), which was the only ratio exceeding the
479 value of 2 out of the 21 ratios calculated (Table S6). Even in this case, the difference
480 between endpoint estimates remains small enough to cause no concern from the risk
481 assessment point of view, as a safety factor of 10 is systematically applied on endpoints
482 from chronic toxicity tests with invertebrates in the EU [17]. The gain following a 56-
483 days test duration (resp. counting eggs) is negligible compared to a 28 days test
484 duration (resp. counting only clutches). This gain is too small to justify the investment in
485 terms of human resources and experimental costs that occur when doubling the
486 experiment duration and significantly increase the workload when counting eggs (which
487 is the most time-consuming part of the experiment). Shorter test duration also reduces
488 risk of failure, both in achieving validity criteria and in issues with equipment [1]. It can
489 be therefore concluded that the optimized test design provides an adequate balance
490 between endpoint accuracy and testing effort.

Conclusion

The present work demonstrated the feasibility, robustness and reproducibility of the experimental protocol designed for testing reproductive toxicity of chemicals with *L. stagnalis* according to the draft OECD Test Guideline. In addition, it allowed optimizing the experimental design in terms of test duration and choice of the core reproductive endpoint. Based on our results a test duration of 28 days is recommended for the reproduction toxicity test with *L. stagnalis*. As the core test endpoint, we recommend to use the mean cumulative number of clutches per individual-day, calculated over 28 days, providing that the number of clutches is determined at least twice a week in six replicates of five snails (at test initiation) per treatment (at least five concentrations) and control. Such a test design was proved as optimal, making the reproduction toxicity test both sensitive and cost effective for estimating accurate ECx values according to current OECD requirements.

Acknowledgements

The authors warmly thank all experimenters for their valuable contribution in collecting the data, in particular Barroso C., Coke M., Collinet M., Dennis N., DeSaeyer N., Handlos F., Kauf A., Kinnberg K.L., Kuhl K., Loureiro S., Lutter M., Örn S., Reategui E., Ruppert R and Salice C. The authors also express their gratitude to Delignette-Muller M.L., Ruiz P. and Veber P. for developing the 'morse' R package and the MOSAIC platform, Charret Q. for helping in writing R codes and Adam C. for the analysis of TBT. Many thanks to Teel C. for her participation in statistical analyses. This study was financially supported by Danish EPA (DK), DEFRA (UK), INRA (FR), ONEMA (FR), and UBA (DE), as well as by

internal resources from the laboratories that took part in the prevalidation and the validation ring-tests.

References

1. OECD. 2010. Detailed review paper on mollusc life-cycle toxixity testing. *Environ. Heal. Saf. Publ. Ser. Test. Assess. - N°121.*, p 182.
2. Ducrot V, Askem C, Azam D, Brettschneider D, Brown R, Charles S, Coke M, Collinet M, Delignette-Muller M-L, Forfait-Dubuc C, Holbech H, Hutchinson T, Jach A, Kinnberg KL, Lacoste C, Le Page G, Matthiessen P, Oehlmann J, Rice L, Roberts E, Ruppert K, Davis JE, Veauvy C, Weltje L, Wortham R, Lagadic L. 2014. Development and validation of an OECD reproductive toxicity test guideline with the pond snail Lymnaea stagnalis (Mollusca, Gastropoda). *Regul. Toxicol. Pharmacol.* 70:605–614.
3. OECD. 2000. Guidance document on aquatic toxicity testing of difficult substances and mixtures. *Environ. Heal. Saf. Publ. Ser. Test. Assess. - N°23.*, p 53.
4. Giusti A, Barsi A, Dugué M, Collinet M, Thomé J-P, Joaquim-Justo C, Roig B, Lagadic L, Ducrot V. 2013. Reproductive impacts of tributyltin (TBT) and triphenyltin (TPT) in the hermaphroditic freshwater gastropod Lymnaea stagnalis. *Environ. Toxicol. Chem.* 32:1552–60.
5. SANCO/12571/2013. 2014. Guidance document on analytical quality control and validation procedures for pesticide residues analysis in food and feed. *Eur. Comm. - Heal. Consum. Prot. Dir.*:1–46.
6. R Core Team. 2015. R: A Language and Environment for Statistical Computing. Available from https://www.r-project.org.
7. Seshan VE. 2015. clinfun: Clinical Trial Design and Data Analysis Functions.
8. Delignette-Muller M., Ruiz P, Charles S, Duchemin W, Lopes C, Kon Kam King G. 2015. morse: Modelling Tools for Reproduction and Survival Data in Ecotoxicology.

9. Delignette-Muller ML, Lopes C, Veber P, Charles S. 2014. Statistical handling of reproduction data for exposure-response modeling. *Environ. Sci. Technol.* 48:7544–51.
10. MOSAIC. 2015. MOdeling and StAtistical tools for ecotoxICology. Available from http://pbil.univ-lyon1.fr/software/mosaic/reproduction/.
11. Ritz C, Streibig J. 2005. Bioassay analysis using R. *J. Stat. Softw.* 12:1–22.
12. Plummer M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proc. 3rd Int. Work. Distrib. Stat. Comput. March*.:20–30. doi:10.1.1.13.3406.
13. Delignette-Muller ML, Dutang C. 2015. fitdistrplus : An R Package for Fitting Distributions. *J. Stat. Softw.* 64:1–34.
14. Sprague JB. 1995. Factors that modify toxicity. In Rand, G.M., ed., *Fundam. Aquat. Toxicol.*, pp 124–163.
15. OECD. 2004. Test No. 202: Daphnia sp. acute immobilisation test. *OECD Guidel. Test. Chem. Sect. 2 – Eff. Biot. Syst.*, p 12.
16. OECD. 2011. Test No. 235: Chironomus sp., acute immobilisation test. *OECD Guidel. Test. Chem. Sect. 2 – Eff. Biot. Syst.*, p 17.
17. Efsa. 2013. Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters. *EFSA J.* 11:267.

Figure legends

Figure 1. Cadmium median EC_{50} estimates from clutch data at day 28 (in red) or at day 56 (in orange), and from egg data at day 28 (in dark green) or at day 56 (in light green). Dotted lines separate laboratories, while the black solid line separates the prevalidation from the validation ring-test.

Figure 2. EC_{50} estimates (medians and 95% credible intervals) as a function of exposure duration (in days) for all laboratories and both endpoints of the prevalidation ring-test. Open symbols indicate the first exposure duration at which the EC_{50} median obtained in a given laboratory becomes similar to that of other laboratories (grey band, which represents the standard deviation of the $EC_{50\text{-}56d}$ for all laboratories from the prevalidation ring-test).

Figure 3. EC_{50} estimates (medians and 95% credible intervals) as a function of exposure duration (in days) for all laboratories and both endpoints of the validation ring-test. Open symbols indicate the first exposure duration at which the EC_{50} value obtained in a given laboratory becomes similar to that of other laboratories (grey band, which represents the standard deviation of the median $EC_{50\text{-}56d}$ for all laboratories from the validation ring-test).