

2016-11

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

Charles, S

<http://hdl.handle.net/10026.1/5257>

10.1016/j.yrtph.2016.07.012

Regulatory toxicology and pharmacology : RTP

Elsevier BV

All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.

1 This is an accepted manuscript of an article published by Elsevier in Regulatory Toxicology & Pharmacology,
2 25 July 2016. Available at DOI: 10.1016/j.yrtph.2016.07.012

3

4

5 **Running head:**

6 Optimizing reproduction toxicity test for aquatic molluscs

7

8 **Corresponding author:**

9 Professor Sandrine CHARLES

10 Laboratoire de Biométrie - Biologie Evolutive

11 Université de Lyon; Université Lyon 1;

12 CNRS; UMR 5558;

13 Bâtiment Gregor Mendel, Mezzanine

14 43 boulevard du 11 novembre 1918

15 F-69622 Villeurbanne Cedex, France

16 Tel. +33 (0)4 7243 2900

17 Mail: sandrine.charles@univ-lyon1.fr

18

19 **Optimizing the design of a reproduction toxicity test with the pond**

20 **snail *Lymnaea stagnalis***

21 **Authors**

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

29 **Affiliations**

30 † Univ Lyon, Université Lyon 1, UMR CNRS 5558, Laboratoire de Biométrie et
31 Biologie Évolutive, F-69100 Villeurbanne, France
32 ‡ Institut National de la Recherche Agronomique (INRA), Centre de Recherche
33 de Rennes, 65 rue de Saint-Brieuc, F-35042 Rennes, France
34 § Bayer Aktiengesellschaft, Crop Science Division, BCS AG-R&D-D-EnSa-ETX-AQ,
35 Alfred-Nobel Straße 50, D-40789 Monheim am Rhein, Germany
36 || The Food and Environment Research Agency (now Fera Science Ltd). Sand
37 Hutton, York, YO41 1LZ, United Kingdom.
38 # Goethe University Frankfurt am Main, Department Aquatic Ecotoxicology,
39 Max-von-Laue-Straße 13, D-60438 Frankfurt, Germany
40 †† Laboratory of Environmental Toxicology and Aquatic Ecology, Faculty of
41 Bioscience Engineering, Ghent University, Belgium

42 ‡‡ Department of Biology & CESAM, Centre for Environmental and Marine
43 Studies, University of Aveiro, 3810-193 Aveiro, Portugal
44 §§ DuPont, PO Box 60, 1090 Elkton Road, DuPont Stine-Haskell Research Center,
45 S315/1369, Newark, Delaware, USA.
46 || || Department of Biology, University of Southern Denmark, Campusvej 55, 5230
47 Odense M, Denmark
48 ## School of Biological Sciences, University of Plymouth, Plymouth PL4 8AA
49 United Kingdom
50 ††† Departamento de Biologia, Universidade de Aveiro, 3810-193 Aveiro, Portugal
51 ‡‡‡ Old School House, Brow Edge, Backbarrow, Ulverston, Cumbria LA128QX,
52 United Kingdom
53 §§§ Department of Pathology Faculty of Veterinary Science Swedish University of
54 Agricultural Sciences P.O. Box 7028 Uppsala, S-750 07 Sweden
55 || || || Department of Environmental Toxicology, Texas Tech University, Lubbock, TX,
56 USA.
57 ### Ibacon GmbH; Arheilger Weg 17, 64380 Rossdorf, Germany
58 †††† Fraunhofer Institute for Molecular Biology and Applied Ecology, Department
59 of Ecotoxicology, Auf dem Aberg 1, 57392 Schmallenberg, Germany.
60 ‡‡‡‡ University of Liège, Laboratory of Animal Ecology and Ecotoxicity (LEAE-
61 CART), Allée du 6 Août, 11, Sart-Tilman, Belgium
62 §§§§ WIL Research, Department of In vitro and Environmental Toxicology, Ashland,
63 United States
64 || || || || BASF SE, Crop Protection – Ecotoxicology, Speyerer Straße 2, D-67117
65 Limburgerhof, Germany

66 **Abstract**

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

84 **Keywords**

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

86

87

88 Introduction

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

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

113 The validation ring-test was conducted from October 2013 to October 2014 according to
114 the draft standard operating procedure. In total, 13 laboratories from academia,
115 government, industry and consultancy, in Europe and North-America, participated in
116 collecting raw data and water samples for statistical and chemical analyses, respectively.
117 Six laboratories (all new compared to the laboratories involved in the prevalidation
118 ring-test) were in charge of testing cadmium (Cd), which had been used in the
119 prevalidation ring-test [2]. Five laboratories tested tributyltin (TBT), four laboratories
120 tested prochloraz (PRO), and two laboratories tested trenbolone (TRB). The choice of
121 these substances was based upon recommendations from the OECD VMG-Eco (Table 1).
122 They were assumed to cause adverse effects on snail reproduction (as confirmed in pre-
123 tests that were conducted for all chemicals except trenbolone). These substances reflect
124 different levels of complexity in terms of toxicity testing; Cd is an “easy-to-test”
125 substance, whereas TBT, PRO and TRB are more difficult substances to test (e.g., use of
126 solvent required for TBT; limited stability of PRO in water, both difficulties being
127 encountered for TRB [3]). Hence, performing the validation ring-test with difficult test
128 substances could contribute to further demonstrate the robustness of the experimental
129 protocol and to identify the most relevant reproduction endpoint in *L. stagnalis*.
130 This paper presents the results of the validation ring-tests for Cd, TBT and PRO, in
131 comparison with those of the prevalidation ring test where applicable. Exposure of
132 snails to TRB up to a mean measured concentration of 776 ng.L⁻¹ had no effects on their
133 reproduction; the corresponding results are thus not presented in this paper. For the
134 remaining substances, ECx were estimated for each laboratory and then compared in
135 order to assess their reproducibility between laboratories. We also investigated the
136 consequence of reducing the exposure duration on both ECx median value and
137 uncertainty. Finally, after having confirmed the low between-laboratory variability

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

141 **Materials and Methods**

142 *Implementation of the validation ring-test*

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

149 *Principle of the reproduction toxicity test*

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

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

164 *Tested chemicals and exposure water sampling and analysis*

165 Specifications of the test chemicals are provided in Table 1. Nominal concentrations for
166 Cd were chosen based on the prevalidation ring-test, namely 25, 50, 100, 200, 400 $\mu\text{g.L}^{-1}$.
167 Nominal concentrations were 87.5, 175, 350, 700, 1400 ng.L^{-1} and 10, 32, 100, 320, 1000
168 $\mu\text{g.L}^{-1}$ for TBT and PRO, respectively. Water samples were collected before and after
169 water renewal, at the beginning, mid-term and end of each experiment for the
170 determination of actual exposure concentrations (42 samples per experiment). Actual
171 Cd concentrations in water were measured in 50 mL acidified samples (triplicates) by
172 atomic adsorption spectrometry (limit of detection: 0.8 $\mu\text{g.L}^{-1}$). Actual TBT
173 concentrations in water were measured in triplicate by coupled capillary gas
174 chromatography to mass spectrometry (GC-MS-MS; ITQ100, Thermo Scientific, USA)
175 according to Giusti et al. [4] with slight modifications. The limit of detection (LOD) was 6
176 ng TBT.L^{-1} and the limit of quantification (LOQ) was 18 ng TBT.L^{-1} (concentrations are
177 expressed in ng TBT.L^{-1} : equivalent in ng Sn.L^{-1} can be calculated by dividing these
178 values by a factor 2.44). The mean recovery efficiency was $99\% \pm 18.6\%$ and was in
179 good agreement with requirements of the SANCO guidance document [5]. PRO samples
180 were analysed directly from filtered samples by LC-MS-MS (LOD: 3.9 $\mu\text{g.L}^{-1}$, LOQ: 1.56
181 $\mu\text{g.L}^{-1}$ and mean recovery efficiency: $70\% \pm 6.3\%$).

182 *Statistical modelling of reproduction data*

183 Solvent controls were used as the reference for statistical analysis of the TBT data (all
184 laboratories). We used the Jonckheere-Terpstra hypothesis test as a way to discriminate
185 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} :

$$223 \quad N_{ij} = \text{Poisson}\left(f(u_i; q) \cdot NID_{ij}\right) \quad (1)$$

224 where $f(u_i; q)$ 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:

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

232 where $q = (EC_{50}, d, b)$, 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.

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

$$236 \quad N_{ij} \sim \text{Poisson}(f_{ij} \times NID_{ij}) \quad \text{with} \quad f_{ij} \sim \text{gamma}\left(\frac{f(u_i; q)}{w}, \frac{1}{w}\right) \quad (3)$$

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

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

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

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

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

257 We thus assumed a normal distribution for $\log_{10}(EC_{50})$ centred on the mean of
 258 $\log_{10}(u_{\min})$ and $\log_{10}(u_{\max})$, with the probability that $\log_{10}(EC_{50})$ lies between
 259 $\log_{10}(u_{\min})$ and $\log_{10}(u_{\max})$ equals to 0.95;

- 260 • As d stands for the reproduction output in controls, we set a normal prior
 261 $N(m_d, s_d)$ based on the data themselves:

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

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 • $\log_{10}(b) \sim U(-2, 2)$ a quasi-non-informative prior for the shape parameter;
 267 • $\log_{10}(w) \sim U(-4, 4)$, 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 $N_{\text{reprodadj}} = N_{\text{reprocumul}}/N_{\text{indtime}}$
 275 computed on a replicate basis (results not shown); this alternative analysis provided

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

278 *Datasets*

279 A full statistical analysis was conducted on all available datasets, i.e., datasets from the
280 prevalidation ring-test [2] and datasets from the validation ring-test presented
281 hereafter. Combining ring-tests (prevalidation and validation), endpoints (number of
282 clutches and number of eggs per clutch) and chemicals (Cd, TBT and PRO) from the
283 participating laboratories resulted in a total of 84 datasets to analyse. For each dataset,
284 EC₅₀ values were estimated for cumulative reproduction per individual-day over 56
285 days, expressed via either the number of clutches, or the number of eggs per clutch.

286 *Optimizing the exposure duration*

287 For each of the considered endpoints, the possible reduction in the experiment duration
288 was investigated by comparing the EC₅₀ estimates (median and 95% credible interval)
289 obtained in a given laboratory at time 21, 28, 35, 42 and 49 days with the median EC₅₀
290 value obtained after 56 days (denoted by EC_{50-56d} hereafter) surrounded with the
291 variability between all laboratories. This inter-laboratory variability was calculated as
292 plus or minus the standard deviation (*sd*) of all median EC_{50-56d} values, separately from
293 the pre-validation and validation ring-tests. We considered as optimal the shortest
294 exposure duration that was outside the inter-laboratory variability range. For this
295 shortest exposure duration, the EC_{50-d} estimate for a given laboratory at day *d* was
296 considered as not different from the EC_{50-56d} estimate, meaning that a stable enough EC₅₀
297 estimate had been reached at day *d* already.

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

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

305 *Comparing results from clutch- and egg-based endpoints*

306 For the chosen optimal exposure duration, we investigated whether the EC₅₀ could be
307 accurately estimated based upon the number of clutches alone or whether eggs must be
308 also counted. For that purpose, we compared the posterior probability distributions of
309 EC₅₀ values, as provided by the Bayesian inference method, using clutch and egg data.
310 We used the R package ‘fitdistrplus’ [13] in order to obtain the Cullen and Frey graph.
311 This skewness-kurtosis plot helps to choose the most appropriate distribution among
312 common ones. Given that priors on EC₅₀ were lognormally distributed, we may expect
313 also a lognormal distribution for the posteriors. Once the suitability of the lognormal law
314 for the posteriors was established, we used the following indices to check similarities
315 between posterior distributions of EC₅₀ estimates from clutch and egg data:

- 316 • the 2.5 and 97.5% quantiles (denoted $Q_{2.5EC_{50}}$ and $Q_{97.5EC_{50}}$, respectively) from
317 the EC₅₀ posterior distribution;
- 318 • the mean, standard deviation and coefficient of variation from the fitted
319 lognormal distribution;
- 320 • the uncertainty of EC₅₀ estimates, namely $Q_{\text{extent}} = Q_{2.5EC_{50}} - Q_{97.5EC_{50}}$.

321 **Results**

322 *Validation ring-test results at day 56*

323 *Test validity*

324 Test validity criteria as stated in the consolidated standard operating procedure were
325 achieved in all laboratories: temperature remained within the 20 ± 1 °C range; oxygen
326 saturation did not drop below 60% air saturation value (ASV; 5.4 mg.L⁻¹ at 20°C);
327 mortality did not exceed 20% in control groups by the end of the test; fecundity in the
328 controls was at least 8 egg-clutches per snail at the end of the 56d test. In addition, each
329 laboratory was able to maintain an appropriate water quality: pH was in the 7.0 - 8.5
330 range; conductivity in the 400 – 800 µS.cm⁻¹ range; and water hardness was in the 140 –
331 250 mg.L⁻¹ range.

332 *Measured exposure concentrations*

333 Mean measured concentrations were calculated for each chemical and laboratory as the
334 arithmetic mean of all measured values over the test duration. They were linearly
335 related to the nominal concentration (see SI, Figure S0). Mean measured Cd
336 concentrations (calculated for all participating laboratories) were 19, 35, 70, 149 and
337 300 µg.L⁻¹, which compare to nominal values of 25, 50, 100, 200 and 400 µg.L⁻¹. Mean
338 measured TBT concentrations were 39, 78, 118, 251 and 435 ng.L⁻¹, which compare to
339 nominal values of 87.5, 175, 350, 700 and 1,400 ng.L⁻¹. Mean measured PRO
340 concentrations were 13, 21, 56, 324 and 765 µg.L⁻¹, which compare to nominal values of
341 10, 32, 100, 320 and 1,000 µg.L⁻¹. The mean measured exposure concentration values
342 specific to each laboratory were used for the estimation of ECx values.

343 *Test results*

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

349 *Reproducibility of results between laboratories*

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

353 *Optimizing the experimental design*

354 For Cd, median EC_{50-56d} ± *sd* intervals used to compare EC₅₀ estimates (median and 95%
355 credible interval) at each exposure duration (from 21 to 56 days) are given in Figures 2
356 and 3 for the prevalidation and validation ring-tests, respectively. Results for TBT and
357 PRO are given in Supplementary Information (Figures S3-S5).

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

367 *Test results at day 28*

368 All datasets corresponding to both the prevalidation and validation ring-tests were
369 analysed simultaneously at day 28. As shown in Table S1, both endpoints were
370 significantly altered within the tested concentration range for all laboratories whatever
371 the chemical, except for Lab. 11 with Cd and the clutch-based endpoint (Jonckheere-
372 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-28d} estimates are detailed in SI
374 (Tables S2, S3 and S4). Results show robust EC_{50-28d} 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_{50-28d} 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).

395 **Choosing the main core endpoint**

396 Overlapping boxplots on Figure 1 (resp. Figures S1 and S2) illustrate the similarity
397 between EC₅₀ estimates from clutch and egg-based endpoints at day 28. Figure S6
398 strengthens this result based on the comparison of full posterior distributions of EC₅₀
399 estimates superimposed to prior ones: distributions have similar positive skewness and
400 similar kurtosis; peaks of distributions are also closely located in most cases.
401 From Table 3, we notice that EC₅₀ medians and uncertainty extents (given by Q_{extend}) are
402 very good proxies of mean and standard deviation of the fitted lognormal distribution:
403 $\mu_{EC_{50}} \simeq \text{Median}_{EC_{50}}$ and $\sigma_{EC_{50}} \simeq Q_{\text{extend}} / 4$. The coefficients of variation confirm these results
404 with equal values from clutch- or egg-based endpoint, except in three cases out of 22
405 comparisons (bold numbers in Table 3). EC₅₀ medians from clutches were generally
406 similar to EC₅₀ medians from eggs (Table 3); indeed, both EC₅₀ medians remained similar
407 based on EC₅₀ median ratios close to 1 (except for Lab.13 with Cd in the validation ring-
408 test).

409 **Discussion**

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

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

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

426 For all tested chemicals, results of the reproduction tests were estimated with good
427 precision, i.e., small 95% credible intervals, indicating that the test protocol and method
428 used to estimate the EC_x values were robust. Results were also homogenous between
429 laboratories, since most of the laboratories provided comparable EC₁₀ (see Table S2-S4)
430 and EC₅₀ values with overlapping 95% credible intervals (Figure 1). For Cd and TBT
431 (with the exception of Lab. 08), a 2-fold difference was obtained between the lowest and
432 the highest estimated EC_{50-56d} values (using either the number of clutches, or the
433 number of eggs per individual-day). For Cd, lower EC_{50-56d} values were found for both
434 endpoints in Lab. 02. The softness of test water used in this laboratory (< 50 mg.L⁻¹ of
435 CaCO₃) may explain this result, as water softness is known to increase the Cd toxicity
436 [14]. A similar trend was already observed in the prevalidation ring-test (see Lab. 07
437 Figure 1, [2]). The high coefficient of variation value for the clutch-based endpoint with
438 Cd in the validation ring-test (52.5%) was due to the high estimate of EC_{50-56d} for Lab. 11
439 (Figure 1). For PRO, inter-laboratory variability in EC₅₀ values was below a factor 2.
440 These results attest to a good reproducibility of the EC_{50-56d} values between laboratories.
441 Indeed, these differences are in the range of acceptable variation defined for reference
442 chemicals in OECD guidelines for acute toxicity tests with invertebrates (i.e., factor 3.5
443 for K₂Cr₂O₇ 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 EC₅₀ estimate. To further strengthen this
470 assumption, we calculated the ratio between EC₅₀ 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 EC₅₀ estimate obtained with
476 the optimized design was 1.3 fold lower than the median EC₅₀ 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.

491 **Conclusion**

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

504

505 **Acknowledgements**

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

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

516

517 **References**

- 518 1. OECD. 2010. Detailed review paper on mollusc life-cycle toxicity testing. *Environ. Heal.*
519 *Saf. Publ. Ser. Test. Assess.* - N°121., p 182.
- 520 2. Ducrot V, Askem C, Azam D, Brettschneider D, Brown R, Charles S, Coke M, Collinet M,
521 Delignette-Muller M-L, Forfait-Dubuc C, Holbech H, Hutchinson T, Jach A, Kinnberg KL,
522 Lacoste C, Le Page G, Matthiessen P, Oehlmann J, Rice L, Roberts E, Ruppert K, Davis JE,
523 Veauvy C, Weltje L, Wortham R, Lagadic L. 2014. Development and validation of an OECD
524 reproductive toxicity test guideline with the pond snail *Lymnaea stagnalis* (Mollusca,
525 Gastropoda). *Regul. Toxicol. Pharmacol.* 70:605–614.
- 526 3. OECD. 2000. Guidance document on aquatic toxicity testing of difficult substances and
527 mixtures. *Environ. Heal. Saf. Publ. Ser. Test. Assess.* - N°23., p 53.
- 528 4. Giusti A, Barsi A, Dugué M, Collinet M, Thomé J-P, Joaquim-Justo C, Roig B, Lagadic L,
529 Ducrot V. 2013. Reproductive impacts of tributyltin (TBT) and triphenyltin (TPT) in the
530 hermaphroditic freshwater gastropod *Lymnaea stagnalis*. *Environ. Toxicol. Chem.*
531 32:1552–60.
- 532 5. SANCO/12571/2013. 2014. Guidance document on analytical quality control and
533 validation procedures for pesticide residues analysis in food and feed. *Eur. Comm. - Heal.*
534 *Consum. Prot. Dir.:*1–46.
- 535 6. R Core Team. 2015. R: A Language and Environment for Statistical Computing. Available
536 from <https://www.r-project.org>.
- 537 7. Seshan VE. 2015. clinfun: Clinical Trial Design and Data Analysis Functions.
- 538 8. Delignette-Muller M., Ruiz P, Charles S, Duchemin W, Lopes C, Kon Kam King G. 2015.
539 morse: Modelling Tools for Reproduction and Survival Data in Ecotoxicology.

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

557

558

559 **Figure legends**

560 **Figure 1.** Cadmium median EC_{50} estimates from clutch data at day 28 (in red) or at day
561 56 (in orange), and from egg data at day 28 (in dark green) or at day 56 (in light green).
562 Dotted lines separate laboratories, while the black solid line separates the prevalidation
563 from the validation ring-test.

564

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

571

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