18-00379

Environmental Toxicology

Comparison of False-Positive Rates of 2 Hypothesis-Test Approaches in Relation to Laboratory Toxicity Test Performance

John F. Fox,[a] Debra L. Denton,[b,]* Jerry Diamond,[c] and Robyn Stuber[b]

[a]Office of Research and Development, US Environmental Protection Agency, Washington, DC

[b]US Environmental Protection Agency, Region 9, Sacramento, California

[c]Tetra Tech, Owings Mills, Maryland, USA

(Submitted 10 July 2018; Returned for Revision 18 August 2018; Accepted 18 December 2018)

Abstract: We compared 2 statistical hypothesis-test approaches (no-observed-effect concentration [NOEC] and test of significant toxicity [TST]) to determine the influence of laboratory test performance on the false-positive error rate using the US Environmental Protection Agency's *Ceriodaphnia dubia* reproduction whole-effluent toxicity (WET) test endpoint. Simulation and power calculations were used to determine error rates based on observed control coefficients of variation (*CV*) for 8 laboratories over a range of effect levels. Average *C. dubia* control reproduction among laboratories was 20 to 40 offspring per female, and the 75th percentile *CV* was 0.10 to 0.31, reflecting a range in laboratory performance. The 2 approaches behave similarly for *CV*s of 0.2 to 0.3. At effects <10%, as *CV* decreases, TST is less likely to declare toxicity and NOEC is more likely to do so. Laboratory performance affects whether a sample is declared toxic and influences the probability of false-positive (and -negative) error rates using either approach. At the 75th percentile control *CV* observed for each

laboratory, 4 laboratories would achieve approximately a 5% false-positive rate using 13 or fewer replicates for this test method. For the remaining 4 laboratories, more replicates would be needed to achieve a 5% false-positive rate. The present analyses demonstrate how false-positive rates are influenced by laboratory performance and WET test design.

Keywords: Whole-effluent toxicity, Bioequivalence, No-observed-effect concentration, *Ceriodaphnia*, Test of significant toxicity

This article contains online-only Supplemental Data.

* Address correspondence to Denton.Debra@epa.gov

Published online XXXX 2019 in Wiley Online Library (www.wileyonlinelibrary.com).

DOI: 10.1002/etc.4347

INTRODUCTION

Whole-effluent toxicity (WET) tests (also referred to as “direct toxicity assessments” or “whole-effluent assessments”) are used by many countries to assess the water quality of treated wastewater effluents and ensure that aquatic life is protected from toxicity of effluents (US Environmental Protection Agency 2002a; Scroggins et al. 2002; Power and Boumfrey 2004; UK Environment Agency 2005; OSPAR Commission 2007). The biological data from these tests (e.g., number of organisms surviving, number of offspring produced, biomass, number of embryos which develop normally) are evaluated using one of several statistical approaches to derive a statistical endpoint estimate for toxicity. That endpoint is then compared with a threshold or criterion to determine whether the effluent meets regulatory standards.

In the United States, the statistical endpoint for permitted effluents is evaluated at a critical effluent dilution (receiving water concentration [RWC]) identified in the facility’s effluent discharge permit. If the effluent is not toxic at the RWC, the effluent is in compliance

with the regulatory permit. Reliance in the United States on a critical effluent concentration for determining compliance has led to the use of hypothesis-testing approaches that determine whether the biological response at the RWC is significantly different (i.e., poorer) from the test control response. Point estimate endpoints (e.g., 25% inhibitory concentration [IC25]) are also used in the United States and elsewhere for effluent toxicity compliance determinations.

Two statistical approaches based on one-sided hypothesis tests are used in the United States to interpret data on survival and sublethal effects in WET tests. These are the no-observed-effect concentration (NOEC) approach, based on traditional null hypothesis significance testing, and a bioequivalence hypothesis-testing approach called the "test of significant toxicity" (TST; US Environmental Protection Agency 2010). The NOEC is an accepted statistical option in Europe (Organisation for Economic Co-operation and Development 2006), the United Kingdom (UK Environment Agency 2007), Canada (Environment Canada 2005), and other countries for effluent toxicity assessments.

The NOEC and other hypothesis approaches (e.g., TST approach) can lead to decision errors (Warren-Hicks and Parkhurst 1996; US Environmental Protection Agency 2000, 2010), commonly termed "false positives" (type I—error of rejecting the null hypothesis when it should be accepted) and "false negatives" (type II—error of accepting the null hypothesis when it should be rejected). Decisions based on point estimates (e.g., IC25) are also subject to false positives associated with imprecision, as reflected in the confidence intervals around the endpoint of interest; we did not address point estimates in the present study. Although numerous publications have discussed sources of laboratory and method variability associated with WET biological endpoints (e.g., Warren-Hicks and Parkhurst 1996; Burton et al. 1996; Chapman et al. 1996; Warren-Hicks et al. 2000), few studies have explored false-positive and -negative rates of

WET endpoints used in regulatory compliance. Furthermore, the few that have made the attempt used very small sample sizes (Diamond et al. 2008) or WET tests using methods that have since been refined (Moore et al. 2000); moreover, laboratory performance has improved since these other studies were conducted (Denton et al. 2011).

Hypothesis-test approaches explicitly incorporate a false-positive rate determined a priori by the investigator (or, in the case of WET compliance, by the regulatory WET program). The false-negative rate then depends on the effect size, number of replicates, and intratest variance, unless it is also determined a priori. Detailed explanations of hypothesis tests such as TST and NOEC and their associated statistical error rates can be found in other sources (e.g., Chapman et al. 1996; Denton et al. 2011; Hothorn 2014). Except for the bioequivalence TST approach, to our knowledge hypothesis tests used in the United States and elsewhere for effluent regulatory compliance explicitly considered only false positives and not false negatives. A percentage of minimum significant difference criterion can be used to indirectly limit type-II errors for the NOEC approach (US Environmental Protection Agency 2002b).

The NOEC and TST approaches have been compared using effluent and ambient toxicity test results for several WET methods (Diamond et al. 2013). However, that observational study did not involve known values for the true percentage effect. Probabilities of declaring toxicity for known values of percentage effect, using the TST and NOEC, were reported for many WET methods (US Environmental Protection Agency 2010). That study demonstrated that method variability (larger control coefficient of variation [*CV*] for the biological endpoint measurement) and replication have an important influence on the probability of declaring toxicity. The US Environmental Protection Agency (2010) study used a limited range of values for true percentage effect, *CV*, and number of replicates. It also reported percentiles of *CV* achieved by a

composite, national sample of laboratories and not for individual laboratories. Although the previous work helps relate the probability of declaring toxicity to the *CV*s being achieved in aggregate, it is important to understand how laboratories differ in endpoint precision (within-test variability), which influences the probabilities of false positives and negatives.

Based on questions raised regarding the false-positive rate of bioequivalence-based approaches such as TST compared to the NOEC approach (California State Water Resources Control Board 2011), the authors have observed a general misunderstanding regarding how such error rates are established in relation to WET method test designs. Thus, a more detailed examination of the TST and NOEC approaches is warranted, to convey more clearly the relation between within-test variability (also called "intratest" or "among-replicate" variability), number of replicates, and probability of declaring toxicity. Also, laboratory performance for many WET methods has improved since they were first introduced in the late 1980s (US Environmental Protection Agency 2010; Denton et al. 2011); thus, more recent data on WET test method variability can provide a more current indication of laboratory performance and how laboratories differ with respect to test performance.

The objectives of the present study were 1) to examine laboratory performance in terms of mean and within-test variability (standard deviation) for the reproduction endpoint of the *Ceriodaphnia dubia* test (US Environmental Protection Agency 2002b) and 2) to evaluate the probability of declaring samples toxic using the TST and NOEC approaches in relation to within-test variability, for a meaningful range of percentage effect values and numbers of replicates. We did not evaluate point estimate approaches because multiple models are in use and various concentration–response patterns need to be examined, requiring a larger, more complex study. The *C. dubia* test was selected because it is frequently required or being considered as a

requirement of freshwater dischargers in the United States, Canada, Japan, and Europe and is commonly used to assess toxicity of freshwater streams in California, where the present study was focused (Anderson et al. 2010). The *C. dubia* reproduction endpoint reflects both survival and reproduction (i.e., average number of offspring produced by all females, including those dying without producing offspring) and is often more sensitive than the survival endpoint. Using results from different laboratories, we demonstrate the influence of laboratory performance on the probability of declaring toxicity and the importance of test design for this WET method in helping to address laboratory performance.

BACKGROUND

The NOEC and TST statistical approaches (US Environmental Protection Agency 2002b, 2010) are applied to sample statistics for the biological endpoints of a WET method, such as reproduction in the *C. dubia* test. The present study is concerned only with sublethal endpoints in short-term WET tests for chronic toxicity.

The NOEC approach is described in US Environmental Protection Agency (2002b) and employs data from multiple effluent concentrations (dilutions) in a WET test, for example, 6.25, 12.5, 25, 50, and 100% of whole effluent. One of the effluent concentrations is the RWC, identified in a discharge permit as the concentration at which no toxicity should occur (i.e., this concentration and more dilute effluent concentrations should not differ significantly from the control). The NOEC statistical approach compares the control to each of the 5 effluent concentrations using a multiple comparisons procedure. The nulhypothesis is that all of the means for effluent are equal to the control mean ($\mu_c = \mu_i$, i = 1, 2, 3, 4, 5); the alternative (one-sided) hypothesis is that at least one of the 5 effluent concentrations has a smaller mean than the control.

The TST statistical approach employs 2 concentrations, an effluent concentration (RWC) and a control (US Environmental Protection Agency 2010; Denton et al. 2011). The null hypothesis assumes that the RWC (parametric) mean is ≤75% of the control mean ($\mu_{RWC} \leq 0.75\mu_C$). The alternative (one-sided) hypothesis is that the RWC mean is >75% of the control mean.

No-observed-effect concentration is a proof-of-hazard statistical test (Hothorn and Hasler 2008; Hothorn 2014); the null hypothesis assumes that the effluent is not toxic; to declare an effluent toxic, there must be a statistically significant difference between the control and the RWC or a lower concentration (a significant difference with a higher concentration does not result in declaring the effluent toxic). The TST is a noninferiority or proof-of-safety statistical test (Parkhurst 2001; Hothorn and Hasler 2008; Hothorn 2014); the null hypothesis assumes that the effluent is toxic; to declare an effluent nontoxic or "safe," the null hypothesis must be rejected.

For NOEC, the type-I error rate (probability of erroneously rejecting the null hypothesis) is set at 0.05; this is a family-wise error rate (i.e., in this case, it pertains to 5 comparisons with control). For TST, the type-I error rate is set between 0.05 and 0.25, choosing the lowest value that would make the probability ≤0.05 of not rejecting the null hypothesis (i.e., declaring a sample "toxic") when the true effect is ≤10%, taking into account the variability and minimum required replication for each WET test. This is a comparison-wise rate, pertaining to a single comparison with control.

For both the NOEC and TST, lower within-test variability and greater replication make it easier to reject the null hypothesis when it is truly false (increasing the power of the test, $\pi = 1 - \beta$). For the NOEC, this means declaring an effluent safe when it should be considered toxic; for

the TST, it means declaring an effluent toxic when it should be considered safe. It is a type-II error to "accept" (with probability β) the null hypothesis when it is truly false. For the NOEC, the type-II error rate β is not set explicitly but depends implicitly on the WET test design (e.g., number of replicates and organisms per replicate) and on the variability of the sublethal endpoint measurement achieved by the laboratory conducting the WET test. For the TST, the type-II error rate is set explicitly at $\beta = 0.05$ when ($\mu_{RWC} \geq 0.90\mu_C$).

The preceding description identifies 3 key effect sizes: no difference in percentage effect (in the NOEC null hypothesis), a 25% difference in means (in the TST null hypothesis), and a 10% difference (for the TST type-II error rate). Thus, 0, 10, and 25% effects (in terms of the parametric means μ_{RWC} and μ_C) are used as benchmarks for evaluating the probability of declaring toxicity in the present study. The term "percentage effect" describes the percentage reduction in the biological measurement. This can refer either to the sample (observed) means or to the parameter values. For example, if *C. dubia* reproduction has a mean of 25 in the control and 20 in the RWC, the percentage effect is **<ZAQ;1>**$[100 \times (25 - 20)]/25$, or 20%.

In practical terms, the definitions of error rates for the NOEC are the reverse of those for the TST. Practical interest centers on the probability of declaring an effluent toxic given its true status. It is simpler to describe outcomes in terms of the probability of declaring an effluent toxic in relation to the effect size (i.e., percentage effect, the relative difference between the true [parametric] mean responses for control and RWC). The present results are presented as probabilities of declaring toxicity in relation to the true percentage effect.

For the present study, we consider that a false positive occurs when an effluent is incorrectly declared toxic when the true effect level is at or below a value deemed to be "acceptable" (i.e., true percentage effect at RWC is ≤10%; $\mu_{RWC} \geq 0.90\mu_C$). A false negative occurs

when an effluent is not declared toxic when the true percentage effect is at a value deemed "unacceptable" (e.g., true percentage effect at RWC is ≥25%; $\mu_{RWC} \leq 0.75\mu_C$).

METHODS

Laboratory test data

Table 1 reports the number of *C. dubia* reproduction WET tests evaluated for each laboratory in the present study. These data were obtained from facilities in California that submit data for regulatory decisions; data are certified as correct under the US National Pollutant Discharge Elimination System (NPDES). Two sources of data were compiled into a single data set for analysis. The first source ("test drive" in the present study, reported in Diamond et al. 2013) consisted of 64 effluent tests conducted before 2012 by 3 facilities using 3 laboratories (the data set originally consisted of 209 tests; we excluded 145 tests for which the laboratory was not known—126 of these were tests of storm water, not effluents). The second source (California Integrated Water Quality System) consisted of 180 effluent tests extracted from California's NPDES compliance database that were conducted during 2012 to 2015 (except one test in November 2011) by 10 facilities using 6 laboratories. One of these laboratories ("D") also occurred in the earlier test drive data set; labels "D1" and "D2" distinguish tests conducted by laboratory D before and after 2012, respectively. This laboratory instituted additional laboratory quality assurance practices after 2012; the mean and *CV* differed between those 2 different time periods, so they are reported separately. Thus, test data were produced by a total of 8 laboratories ("A"–"H"). All *C. dubia* effluent tests were conducted using one organism in each of 10 replicates for the control and 5 effluent concentrations in accordance with the minimum test design required in this WET method (US Environmental Protection Agency 2002b). Dilution

water used in these tests was generally “moderately hard water,” although some tests were conducted using “hard” or “very hard water” (US Environmental Protection Agency 2002b).

Accurate transcription of original data was verified by 2 analysts working independently. Data were entered and verified using spreadsheets. Data manipulation and analysis were conducted using the R statistical programming system (R Development Core Team 2016). Data imported into R from spreadsheets were checked against the spreadsheet using frequencies, ranges, and totals and by detailed visual inspection. Replicate test data were organized by laboratory and test identification number for subsequent analysis.

Calculating the probability of declaring toxicity

The probability of declaring toxicity was evaluated in 3 ways for the TST: 1) by an analytic (mathematical) power calculation using the noncentral *t*-distribution, 2) by simulations, and 3) by a resampling approach that avoided parametric assumptions. For the NOEC, only the simulation approach was used, for reasons noted later.**<ZAQ;2>**

The *C. dubia* reproduction endpoint can be represented as a normally distributed variate, and its means are adequately represented by the *t*-distribution (Zheng et al. 2013; US Environmental Protection Agency 2010). Mean responses (normalized by standard deviation [SD]) for each concentration are well represented by the *t*-distribution. These means and SDs are the basis of the TST and NOEC calculations. We considered it prudent to evaluate the extent of agreement between the properties of the laboratory test data and the statistical assumptions (normality and homogeneity of variances) used for our calculations. Laboratory data were examined using normal probability plots and the Shapiro-Wilk test for normality. The relations between SDs and means were examined for controls and for all tested concentrations (Supplemental Data, Part 2).

Power calculations for the TST

The probability of declaring samples toxic is readily calculated directly for the TST using statistical functions in R for the noncentral *t*-distribution. Thus, the probability of declaring toxicity (by failing to reject the null hypothesis for the TST; US Environmental Protection Agency 2010) can be calculated for specified values of percentage effect, number of replicates, and SD or *CV* (Supplemental Data, Part 4). We also calculated the number of replicates necessary to achieve a probability 0.05 of declaring toxicity at 10% effect given the observed *CV* for each laboratory.

Simulations for the TST and NOEC

A direct calculation of the probability of declaring toxicity was not possible for the NOEC approach. Simulation of a multiconcentration WET test was necessary because NOEC evaluation requires multiple decisions in a flowchart, calculations of Dunnett's or Steel's test, and application of percentage of minimum significant different (PMSD) bounds (US Environmental Protection Agency 2002b). Simulations were therefore used to estimate the probability of declaring toxicity using the NOEC approach. The same simulated data (using only the control and 100% effluent concentration) were also used to compute the TST hypothesis test. The simulation results for TST were essentially identical to results from the mathematical power calculation.

The probability of declaring toxicity with the NOEC approach depends on the concentration–response relationship. Thus, a comprehensive evaluation would require simulating multiple scenarios for patterns of response across the range of effluent concentrations. Instead, a simpler case was examined: a control and 5 concentrations with the mean response decreasing linearly (i.e., by constant decrements) from the one concentration to the next (starting at 0%

effluent, the control), until the mean response reached a specific percentage effect (ranging from 0 to 50%) at the highest concentration (100% effluent). The parametric mean response was 25 neonates for the control, which is close to the median for a large sample of tests and laboratories (US Environmental Protection Agency 2010). The 100% concentration was selected as the RWC because 11 of 13 dischargers (and 196/244 tests) in the present study had the RWC set at 100% effluent. In the simulations, the percentage effect parameter at the RWC ranged from 0 to 50%. Setting the RWC at the highest, 100%, concentration is likely to yield a slightly greater probability of declaring toxicity than when the RWC is the middle concentration (for a given percentage effect at the RWC). Each concentration had 10 replicates. Variances of simulated data were homogeneous across concentrations. Data sets ("WET tests") were simulated 10 000 times for each combination of maximum percentage effect (at 100% effluent) and control *CV*. These combinations were maximum percentage effect = 0, 5, 10, 15, 20, 25, 30, 40, and 50% and *CV* = 0.10, 0.20, 0.25, 0.30, and 0.40 (note, these are parameters, not observed values). The sampling error for the estimated probability of declaring toxicity based on simulation is approximately 0.05 ± 0.00427 for a probability of 0.05 and 0.10 ± 0.00588 for a probability of 0.10, calculated as 1.96 standard errors using the normal approximation.

For simulated data sets, the NOEC was determined using the statistical approaches (US Environmental Protection Agency 2002b) for equal numbers of replicates: if tests for normality and homogeneity of variances were not rejected, analysis of variance was followed by Dunnett's test for multiple comparisons to a control; otherwise, Steel's test was used (this occurred for 2% of simulated tests). Simulations were generated using normally distributed data with equal variances and equal replication, which satisfied assumptions for Dunnett's test. The requirement (test acceptability criterion) that the control mean be at least 15 offspring per female for the *C.*

dubia reproduction endpoint was applied. A finding of toxicity occurred if any concentration less than or equal to the RWC (in this case, 100%) differed significantly from the control, with exceptions as required for PMSD bounds (US Environmental Protection Agency 2002b). *Results* and Supplemental Data (Part 4) report how often the test acceptability**<ZAQ;3>** criterion was violated (almost never) and how often the PMSD bounds were applied.

If the PMSD calculated for the simulated data exceeds the upper bound for *Ceriodaphnia* (47%) but no concentration differs significantly from control, a second water sample must be taken within 2 wk and a new WET test conducted. This outcome was rare in the simulations except at $CV = 0.40$, when it accounted for 12 to 19% of tests. The probability of declaring toxicity in the second test should be the same as for the first test if all parameters are unchanged. Therefore, we estimated the overall probability of declaring toxicity using the NOEC by multiplying the probability that the second test is required by the probability that the first test is declared toxic and adding this to the probability of declaring toxicity for the first test.

In these simulations and in mathematical power calculations, the SDs were the same for control and RWC, which implies that the *CV* at the RWC is larger (because the mean is smaller). This exaggerates the variability at the RWC: based on the data from the present study (Supplemental Data, Part 2), the SD calculated from the data is approximately the same across a range of concentrations when the percentage effect is <25%.

Resampling control data

We also evaluated the probability of declaring toxicity for the TST by resampling the laboratory test data, rather than simulating data from a normal distribution. This avoids parametric assumptions about the distribution of the data and can validate the parametric approach if the results agree closely. Control replicates for each of the 244 WET tests in the *C.*

dubia reproduction WET method were resampled 10 000 times. Each time, a pair of samples (each with 10 replicates, sampled with replacement) was taken, comprising the "control" concentration and the "RWC." The data for the RWC were multiplied by 0.9 to induce a 10% effect. This means that the SD for the RWC samples was 0.9 times that for the control (the SD wasroportional to the mean, so the *CV* was the same for control and RWC). "Tests" for which the sample mean for control was fewer than 15 neonates were excluded from the calculation of probability.

This bootstrap resampling approach was designed to reflect the variability among control replicates based on actual laboratory test data collected in the present study. If the data are not exactly described by the normal distribution, this approach should provide a more realistic estimate of the probability of declaring toxicity. If the data are well described by the normal distribution, this approach should agree closely with a mathematical power calculation for the TST that uses the sample statistics (*CV*, mean, and SD) of the laboratory data.

RESULTS

Laboratory differences in variability and reproduction

Table 1 summarizes the sample statistics of control means, *CV*s, and SDs for the 8 laboratories contributing *C. dubia* reproduction data. Differences in average control reproduction among laboratories were apparent, ranging from 20 to 40 offspring per female. The SDs also differed among laboratories, ranging between 2.9 and 7.1. The 75th percentile *CV* for the 8 laboratories ranged between 0.13 and 0.31. Laboratory performance exhibited by the sample of 8 California laboratories is similar to that observed nationally (Table 2; US Environmental Protection Agency 2010); thus, the laboratories represented in the present study are expected to

achieve the same probabilities of declaring toxicity as were reported by the US Environmental Protection Agency (2010) in a national study.

The *C. dubia* reproduction control data generally agreed with assumptions of the unequal-variance *t* test used for the TST. Normal probability plots and the Shapiro-Wilk test (Supplemental Data, Figure S1) indicated that normality is a well-supported assumption except for one laboratory (laboratory A, $n = 43$, $p < 0.001$). Boxplots of control SDs and means revealed large laboratory differences in means and smaller differences in SDs (Supplemental Data, Figures S2 and S3). Control SDs appeared to be constant across the range of control means (Supplemental Data, Figures S4 and S5). Using all concentrations (control and effluent), SDs appeared to increase with the mean offspring per female, up to a mean of approximately 15 (these low values for reproduction involved effluent concentrations, not controls), and then were constant or decreased very slightly with increasing mean (Supplemental Data, Figure S6 and Part 2.3). This observation bears on the methods used for simulations and resampling, which assumed constant SD or constant *CV*, respectively.

Probability of declaring toxicity: Power calculations

Figure 1 shows the probability of declaring toxicity for the TST in relation to *CV* when there is no difference between the control and RWC and when there is a 10% effect, based on resampling. When there is no difference, the probability of declaring toxicity is <5% up to a *CV* of approximately 0.25. When there is a 10% effect, this probability is <5% up to a *CV* of approximately 0.17 and exceeds 10% when the *CV* exceeds approximately 0.20. In resampling, the proportion of "tests" failing the test acceptance criterion for the control mean (it must be at least 15) was very low (≤0.003 of "tests") for 7 laboratories, 0.018 for laboratory H, and 0.100 for laboratory A. The result for laboratory A is not surprising because 11 out of 43 control means

were between 15 and 16 neonates per female, the lowest control mean observed among the laboratories in the present study.

Probability of declaring toxicity: Simulations

Figure 2 summarizes the simulation results for the probability of declaring toxicity in *C. dubia* reproduction tests using the TST and NOEC approaches. Even at the highest *CV*, only a small fraction (<0.0013) of tests were rejected because the control mean was <15. This result was expected given the simulation values for the control means and *CV*s, which are typical for experienced laboratories. Good laboratory practices will prevent test rejection and repetition related to failing this test acceptance criterion.

Considering the results of our analyses based on the minimum required number of replicates (n = 10 in Figure 2, bottom row), increasing *CV* decreases the power to declare toxicity at any given percentage effect for both approaches and more so for the NOEC than the TST. As *CV* increases, the TST is increasingly likely to declare toxicity at low percentage effect (also shown in US Environmental Protection Agency 2010). As *CV* decreases, the NOEC is increasingly likely to declare toxicity at a low percentage effect. These outcomes are a direct consequence of the different hypothesis-test approaches used by the NOEC and TST.

Considering the influences of increased replication and decreased *CV* generally, the most precise results are obtained with a *CV* of 0.1 and 30 replicates and the least precise results with a *CV* of 0.4 and the minimum number of replicates (10). Decreasing the *CV* and increasing the number of replicates will increase the precision of the hypothesis-test statistic, although halving the *CV* (from the national average, ~ 0.25) may have a greater influence than doubling the required minimum number of replicates.

The curves for the TST pivot around a point placed at 25% effect and 0.80 probability of declaring toxicity, becoming steeper for smaller *CV* and greater replication. That is a consequence of the TST hypothesis test in which the RWC mean must exceed 75% of the control mean to “pass” at alpha = 0.20 for the *Ceriodaphnia* reproduction endpoint. The curves for the NOEC are “anchored” at 0% effect and probability 0.05 of declaring toxicity; however, at low *CV*, the lower bound PMSD is triggered, recharacterizing significant but small effects as "not significant," so the curve dips below probability 0.05. Thus, the curves for the NOEC also become steeper for smaller *CV* and greater replication but from a pivot point near 0% effect and probability 0.05. This is a consequence of the proof-of-hazard hypothesis-test approach used for the NOEC (which must reject the hypothesis of 0% effect at alpha = 0.05 to show that toxicity is present at some concentration at or below the RWC).

Considering effects of 10% or less, the NOEC is more likely than the TST to declare toxicity for *CV* <0.20 and less likely to do so for higher *CV*s with the minimum required replication. This creates a disincentive to increase test precision (via lower *CV* or higher replication) using the NOEC approach. For 10 replicates and *CV* = 0.3 or 10 to 20 replicates and *CV* = 0.4, the NOEC is less likely to declare toxicity.

Ideally, the probability of declaring toxicity should increase abruptly near the desired percentage effect threshold, having low probability of declaring toxicity for biologically inconsequential effects and high probability of declaring toxicity at effects at or near the threshold for unacceptable toxicity. The TST comes closest to this ideal for *CV* = 0.2 to 0.3 and 20 to 30 replicates.

The results for the NOEC include the possibility that the PMSD upper bound may be exceeded, requiring resampling and a new WET test. This occurred very infrequently except for

the case of *CV* = 0.4 and 10 replicates. In that case, 12 to 19% of tests had NOEC > RWC and PMSD > 47 (the maximum acceptable PMSD for the *Ceriodaphnia* reproduction endpoint using the NOEC approach). Most of the simulated WET tests that triggered repeat sampling (i.e., PMSD > 47 and found "not significant" by the NOEC approach) were declared toxic using the TST (Supplemental Data, Part 4.2). This is evident from the relative pobabilities of declaring toxicity using the TST and NOEC in Figure 2.

Laboratory differences and TST error rates

Figure 3 **<ZAQ;4>**shows the probability of declaring toxicity with the TST at 0 and 10% effect as a function of number of replicates and the control *CV* parameter, when the SD parameter is the same for the control and RWC, based on a mathematical power calculation. A target probability of ≤0.05 for declaring toxicity at a 10% effect using the TST was identified in US Environmental Protection Agency (2010).

At 0% effect (Figure 3), the target probability of 0.05 is achieved with 10 replicates when the *CV* is approximately 0.25 or less. However, approximately 15 to 30 replicates are required as *CV* increases from 0.30 to 0.40 to achieve the same target probability of 0.05. At 10% effect (Figure 3), the probability of declaring toxicity is >0.05 for *CV*s exceeding approximately 0.15. That is approximately a 50th percentile *CV* for this WET method based on a large nationwide sample (Table 2). Doubling the number of replicates (20 replicates per concentration) decreases the long-run probability of declaring samples toxic with a 10% effect, but that probability is >0.05 at *CV*s >0.20. Tripling the number of replicates to 30 would reduce the probability below 0.05 for *CV*s up to approximately 0.25. For 30 replicates and with a *CV* of 0.3 to 0.4, the probability of declaring toxicity for a 10% effect is estimated to be 0.09 to 0.21.

Table 3 shows the expected long-run probability of declaring toxicity at a 10% effect for each laboratory in the present study, based on resampling each laboratory's controls and applying the TST approach, for *C. dubia* reproduction. The influence of higher control *CV*s is again evident. Table 3 also shows the number of replicates needed to achieve a 5% probability of declaring a sample toxic when the true percentage effect is 10% and the true (parametric) control *CV* equals the observed 75th percentile *CV* (based on resampling; the *CV* parameters for control and RWC are the same, so the SD for RWC is 0.9 times the control SD). The observed 75th percentile was chosen to provide a suitable margin of error, but this choice will depend on the number of *CV* observations in any particular case (see Supplemental Data, Part 2.4). More than 30 replicates would be needed to reduce the probability to 0.05 at 10% effect for laboratories having *CV*s >0.25. Laboratories having *CV*s <0.15 can achieve a probability of 0.05 or less at 10% effect using the minimum number of 10 replicates for this WET method reproduction endpoint. In between this range, 10 to 30 replicates would be needed; fewer than 20 replicates are needed when *CV* is <0.21. These findings agree with previous work on TST (US Environmental Protection Agency 2010): the probability of declaring a sample toxic using the TST increases with percentage effect, and this probability depends primarily on the within-test variability and on the number of replicates used in the test.

In Table 3, there is some discordance between the calculation of replicates needed based on a mathematical power calculation using the observed 75th percentile *CV* and the probability of declaring toxicity with 10 replicates based on resampling. This occurred mainly because the probability from resampling tracks the laboratory differences in average *CV*s rather than the 75th percentiles. For example, compare laboratories H and F, which have the same 75th percentile *CV* and thus the same number of replicates needed on that basis; laboratory F has a smaller average

CV and a smaller expected probability of declaring a sample toxic, based on resampling, when the sample has a 10% effect.

Figure 4 shows the probability of declaring toxicity using the TST with 2 WET test endpoints, *Ceriodaphnia* reproduction and red abalone larval development (US Environmental Protection Agency 1995) at the minimum required replication, in relation to percentage effect and 3 values of control *CV*. These curves are based on mathematical power calculations for the TST. Figure 4 shows the influence of precision (low within-test variability) on the probability of TST declaring toxicity at larger effect levels and not declaring toxicity for small effects. The WET test methods that require a greater minimum number of replicates or a greater number of organisms examined per replicate are likely to have lower control *CV*s and, therefore, greater statistical power to reject the null hypothesis. For example, the red abalone larval development test is capable of relatively high precision (control *CV*s are frequently <0.10; US Environmental Protection Agency 2010) because of the greater number of organisms examined in that test method. Using the TST, this test can detect a 25% effect with a probability of 0.95 and has a negligible chance of declaring toxicity for small effects (<10%). Using the NOEC approach, this WET test method is capable of statistically distinguishing very small effluent effects (<5% effect), thereby declaring toxicity because statistical power is relatively high (Diamond et al. 2013).

DISCUSSION

Agreement of resampling and power calculations for the TST

Control data were resampled as a way to avoid parametric assumptions about the distribution of the data. The results agreed well with the mathematical power calculation that assumed normally distributed data and SD proportional to the mean (Figure 1). Mean values of

reproduction data for *C. dubia* controls are well described by a normal distribution (Supplemental Data, Part 2; Zheng et al. 2013). Although the tails of the distribution of means are shortened by the infrequent rejection of means of fewer than 15 offspring per female and the termination of the test (US Environmental Protection Agency 2002b; when 60% or more of surviving females have produced 3 broods), this favors robustness of the *t* test, and variance heterogeneity among concentrations is accommodated successfully by use of the unequal-variance *t* test (Ruxton 2006; US Environmental Protection Agency 2010; Zheng et al. 2013).

Comparison of the TST and NOEC

The 2 approaches are increasingly likely to declare toxicity as percentage effect increases for a given *CV* (Figures 1, 2, and 4), as they were designed to do. For the TST, lower control *CV*s reduce the probability of declaring toxicity for effects <25% and increase that probabiity for effects >25% (this is inherent in the design of the test statistic). For the NOEC, lower control *CV*s increase the probability of finding a significant effect of any size (even small effects). This tendency is mitigated by applying the "lower-bound PMSD," which means that if an estimated effect is significant but <13% (*Ceriodaphnia* reproduction), it is reclassified as not significant. As control *CV*s increase, the NOEC approach has an increasing probability of declaring any positive effect to be nontoxic. It is well known that the NOEC can fail to declare toxicity when within-test variability is high (Chapman et al. 1996; Denton et al. 2003, 2011; Diamond et al. 2011; Landis and Chapman 2011). Increasing replication is qualitatively analogous to decreasing *CV* because it reduces the variance of the sample statistic used in the hypothesis tests for the TST and NOEC. These are well-known consequences of the TST and NOEC approaches.

In comparing the relative chance that the NOEC and TST approaches will declare toxicity across a range of effect sizes, for realistic values of *CV* and replication, the NOEC and

TST differ most at high and low *CV*s across a range of effect levels (Figure 2). This is a direct consequence of the differing null hypotheses. Previous analyses have shown that effluent samples with percentage effect as high as 25 to 30% may not be declared toxic using the NOEC approach at *CV*s corresponding to the 90th percentile for the *C. dubia* reproduction test (Diamond et al. 2013). Inability to detect what is generally considered a significant biological effect (e.g., 25%) is a clear disadvantage for a regulatory compliance program. Also, with the NOEC approach, high test precision results in a high probability of declaring effluent toxicity for small effects that are deemed biologically inconsequential. This behavior of the NOEC is related to using a null hypothesis of no difference, as noted before (Diamond et al. 2013). The bioequivalence test approach used by the TST can address these concerns so long as the laboratory can achieve sufficient within-test precision to minimize the probability of an effluent being declared toxic at effect levels that are acceptable according to regulatory policy.

Variation among laboratories

It is apparent that laboratories having different levels of precision will differ in the probability of declaring toxicity using the TST and the NOEC approaches. In the present study, laboratory A had *CV*s ranging up to 0.30 or higher (Table 1), which could translate to probabilities of declaring toxicity as high as 0.34 for a 10% effect using the TST (Table 3), given the minimum test design of 10 replicates for the *C. dubia* reproduction test. Laboratory A was the only laboratory with demonstrably non-normal data, caused mainly by a high proportion of tests (11 of 43) in which the control means barely met one of the test method acceptability criteria for reproduction (average of 15 or more neonates per female). Laboratories having relatively low *CV*s for this WET test method (≤0.15) are expected to achieve a probability of 0.05 or less with a 10% effect using the TST and 10 to 20 replicates (Table 3). When the

percentage effect is 0%, all of these laboratories would have probabilities of 6% or less for declaring toxicity (3 laboratories exceed 5%) using the minimum test design for the *C. dubia* reproduction endpoint.

Varying precision among laboratories has different consequences for the NOEC and TST given their different null hypotheses (Figure 2). For the *C. dubia* reproduction endpoint, relatively high (but achievable) precision ($CV \leq 0.2$) results in probabilities >0.20 for declaring toxicity at a 10% effect using the NOEC approach, whereas for the TST the probabilities are <0.20. Low test precision ($CV \geq 0.3$) results in a relatively low probability (<0.50) of declaring an effluent as toxic at a 25% effect using the NOEC, whereas the TST maintains probabilities >0.80.

Implications for improving practices

The present analyses point to the need for laboratories to track their control *CV* and, when necessary, adopt measures to decrease within-test variability, thus ensuring quality data for decision-making (Diamond et al. 2008). Variation among laboratories in means and *CV*s for *C. dubia* reproduction results in interlaboratory variability in the probability of declaring toxicity. Using the TST, it is to the permittee's advantage to select laboratories that have smaller control *CV*s. The effect of larger *CV*s on the probability of declaring toxicity at small percentage effect can be counteracted predictably by increasing replication, which decreases the standard error (denominator) used in the *t* test for the TST. Using the NOEC approach, increasing the number of replicates or otherwise increasing within-test precision increases the probability of declaring toxicity across the full range of effect sizes, including small effects (≤10%).

The effect of improved laboratory performance for the *C. dubia* reproduction test can be seen by comparing the estimated probabilities of declaring toxicity for laboratory D over time.

Before refinements in laboratory technician training and additional quality control steps, the mean reproduction and 75th percentile *CV* were 31.6 and 0.31, respectively (n = 30; Table 1). That *CV* translates to approximately a 34% probability of declaring toxicity at 10% effect, using the minimum test design of 10 replicates (Figures 1 and 3). After quality assurance/quality control and training improvements were instituted, the control mean reproduction and 75th percentile *CV* forthis laboratory and test endpoint were 40.0 and 0.17, respectively (n = 57; Table 1). The lower *CV* translates to a probability between 5 and 10% at a 10% effect in the RWC using the minimum required test design (Figures 1 and 3). The probability could be lowered below 5% by increasing replication to 20.

Limitations of the present study

Several choices made for the simulations limit their generality, including the adoption of a linear decrease in responses across the range of concentrations, setting RWC at the highest concentration, and assuming homogeneous variance of responses across the range of effluent concentrations. To evaluate the behavior of the NOEC approach, it was necessary to choose a concentration–response curve and to choose the concentration for the RWC. The NOEC could exhibit different behaviors for different shapes of response curves across the concentration series or when the RWC is an intermediate concentration. As noted, setting the RWC at the highest concentration may slightly increase the chance of declaring toxicity. Although this matter deserves further investigation, the effort required for a thorough study is substantial (e.g., Bailer et al. 2009). The present study was primarily intended to reveal the behavior of the TST approach, which depends on only 2 concentrations (control and RWC). Note that for most permittees in the present study, the RWC is set at 100% effluent, so the present results for the

TST and NOEC should yield a fair assessment of the relative performance of those 2 approaches for such effluent discharges.

For power calculations and simulations, the variance parameter was homogeneous (thus for smaller response means, *CV* is greater). The decision to make the variance of responses homogeneous was based on the data, as noted in the present study (see Supplemental Data, Part 2). Using a constant SD tends to overestimate the false-positive rate (Supplemental Data, Part 4.1). For the resampling analysis (and the curves shown with resampling results in Figure 1), the SD was proportional to the mean: when the mean parameter for RWC is 0.9 × (control mean), the SD parameter for RWC was 0.9 × (control SD). A 10% decrease in the response mean is considered a small effect. Power calculations show that the choice of homogeneous versus proportional SD has little influence over the probability of declaring toxicity for small effects (Supplemental Data, Part 4.1).

The present study used laboratory average or median *CV*s. It is important to note that sample statistics for *CV*, mean, and SD are subject to sampling variation. To apply our methods and inferences, laboratories should consider using upper confidence limits for *CV* or SD estimates, to err on the side of caution (Supplemental Data, Part 2.4).

False positives and false negatives

The regulatory program in the United States has identified a 25% effect as a threshold for unacceptable toxicity, both for the TST hypothesis-test approach and the IC25 point estimate approach (US Environmental Protection Agency 2002b, 2010). For the TST approach, up to a 10% effect at the RWC effluent concentration is deemed acceptable. We note, however, that a 10% effect is not necessarily considered negligible in regulatory programs. For example, the US Environmental Protection Agency's revised selenium water quality criterion is based on a 10%

effect on larval fish development (US Environmental Protection Agency 2016), which is well below the 25% effect used as the basis for IC25 in WET testing in the United States. In addition, toxicity assessments of effluents and chemicals in the European Union are generally based on either 0% (i.e., no difference from control) or a 10% effect level in their chemical registration process (European Commission 2003; OSPAR Commission 2007).

Declaring an effluent toxic when the true percentage effect is zero is clearly a false positive. Extending the definition of false positive to larger effects is inherently a policy decision, as discussed in US Environmental Protection Agency (2010). Obviously, the true percentage effect is unknown for effluent samples or receiving waters. Blank studies could, in principle, be used to estimate a false-positive rate for a given test method; however, a very large number of tests is needed to estimate the rate accurately, and the estimate would be conditional on a laboratory's performance characteristics, such as its consistency in achieving a high control mean and a low control SD, as demonstrated in the present study. It is important to emphasize that declaring a single effluent sample toxic or not toxic, using any statistical approach, cannot be identified unambiguously as a false-positive or false-negative event. Also, samples of unknown toxicity or composition do not provide evidence about the error rates of a WET method or a statistical approach.

False-positive and false-negative error rates are linked to the test design and laboratory performance of the test method and to the effect level identified with these errors. Increased replication along with lower within-test variability will increase test statistical power and decrease error rates, as demonstrated in the present study. The NOEC approach does not directly control false negatives. As demonstrated in the present study, the NOEC approach, using the

minimum required replication, could declare an effluent “safe” with a true effect of 30% or more in the *C. dubia* reproduction test if within-test variability is high.

The results for the *C. dubia* reproduction endpoint are not representative of all WET endpoints and methods. For example, the red abalone (*Haliotis rufescens*) larval shell development WET method (US Environmental Protection Agency 1995), which is commonly required for marine discharges in California and some other Pacific coast states, examines larval shell development for 100 organisms in each of 5 replicates. The CVs for this WET method are an order of magnitude lower than those observed for the *C. dubia* reproduction test (US Environmental Protection Agency 2010). Thus, for the *H. rufescens* and *Mytilus* species WET test methods, even with a percentage effect of 10% at the**<ZAQ;5>** RWC, samples are often declared toxic using the NOEC approach, whereas they are not declared toxic using the TST (Diamond et al. 2013). Similar results are expected for other WET test methods that examine a relatively large number of organisms per replicate and/or more replicates, such as the echinoderm fertilization WET test method (US Environmental Protection Agency 1995). Thus, the WET test design influences the ability to distinguish samples having small and large effects using hypothesis statistical approaches such as the TST and the NOEC.

The bioequivalence null hypothesis encourages a laboratory to achieve higher test power (by minimizing within-test variability and increasing replication) to demonstrate that the effluent is safe and not a potential hazard (McDonald and Erickson 1994; New Mexico Energy, Minerals and Natural Resources Department, Mining and Minerals Division 1999; Hoenig and Heisey 2001; Denton et al. 2011; Hothorn 2014). The statistical literature has long supported applying properly stated null and alternative hypotheses and appropriate test designs (McBride et al. 1993; Erickson and McDonald 1995; Parkhurst 2001; Streiner 2003). The regulatory management

decisions used to define the TST also address the issues inherent in determining statistical significance without defining what is biologically significant (McBride et al. 1993; Parkhurst 2001; Van der Vliet et al. 2012; Hothorn 2014).

Water quality agencies can develop WET methods and evaluate laboratory performance to minimize long-run error rates using test designs that ask and answer the appropriate question statistically (Diamond et al. 2011). Those who use WET methods and engage in discussions about their error rates need to recognize that an individual test result (for a water sample of unknown toxicity) cannot be declared either a "false positive" or a "false negative" because error rates are based on long-run statistical properties.

Supplemental Data—The Supplemental Data are available on the Wiley Online Library at DOI: 10.1002/etc.4347.

Acknowledgment—ICF and Tetra Tech conducted data entry and quality assurance under contract with the USEPA. D. Farrar and R. Erickson of the USEPA and reviewers for this journal provided review and comments that substantially improved the present study.

Disclaimer—The views expressed in the present study are those of the authors and do not necessarily reflect the views or policies of the USEPA.

Data Accessibility—Data, associated metadata, and calculation tools are available on Figshare at DOI: 10.6084/m9.figshare.7122812

REFERENCES

Anderson B, Hunt J, Markiewicz D, Larsen K. 2010. Toxicity in California waters. Surface Water Ambient Monitoring Program. California State Water Resources Control Board, Sacramento, CA, USA.

Bailer AJ, Elmore RT, Shumate BJ, Oris JT. 2009. Simulation study of characteristics of statistical estimators of inhibition concentration. *Environ Toxicol Chem* 19:3068–3073.

Burton G, Arnold R, Ausley L, Black J, DeGraeve GM, Fulk F, Heltshe J, Pelltier W, Pletl J, Rodgers J Jr. 1996. Effluent toxicity variability. In Grothe DR, Dickson KL, Reed-Judkins DK, eds, *Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving System Impacts*. Special SETAC Publication. SETAC, Pensacola, FL, USA, pp 131–156.

California State Water Resources Control Board. 2011. Prevailing comments on the draft policy for toxicity assessment and control. Sacramento, CA, USA. [cited YYYY Month Day]. Available from: https://www.waterboards.ca.gov/water_issues/programs/state_implementation_policy/docs/cmmnts_policy.pdf**<ZAQ;6>**

Chapman GA, Anderson BS, Bailer AJ, Baird RB, Berger R, Burton DT, Denton DL, Goodfellow WL, Heber MA, McDonald LL, Norberg-King TJ, Ruffier PJ. 1996. Methods and appropriate endpoints. In Grothe DR, Dickson KL, Reed-Judkins DK, eds, *Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving System Impacts*. Special SETAC Publication. SETAC, Pensacola, FL, USA, pp 51–130.

Denton DL, Diamond J, Zheng L. 2011. Test of significant toxicity: A statistical application for assessing whether an effluent or site water is truly toxic. *Environ Toxicol Chem* 30:1117–1126.

Denton DL, Fox JF, Fulk FA. 2003. Enhancing toxicity test performance by using a statistical criterion. *Environ Toxicol Chem* 22:2323–2328.

Diamond J, Denton DL, Anderson BA, Phillips B. 2011. It is time for changes in the analysis of whole effluent toxicity data. *Integr Environ Assess Manag* 8:351–358.

Diamond J, Denton DL, Roberts JW Jr, Zheng L. 2013. Evaluation of the test of significant toxicity for determining the toxicity of effluents and ambient water samples. *Environ Toxicol Chem* 32:1101–1108.

Diamond J, Stribling J, Bowersox M, Latimer H. 2008. Application of effluent toxicity as an indicator of aquatic life condition in effluent-dominated streams. A pilot study. *Integr Environ Assess Manag* 4:456–470.

Erickson W, McDonald L. 1995. Tests for bioequivalence of control media and test media in studies of toxicity. *Environ Toxicol Chem* 14:1247–1256.

Environment Canada. 2005. Guidance Document on statistical methods for environmental toxicity tests. Report EPS 1/RM/46. Ottawa, ON, Canada.

European Commission. 2003. Technical Guidance Document on Risk Assessment in support of Commission Directive 93/67/EEC on Risk Assessment for new notified substances, Commission Regulation (EC) No 1488/94 on Risk Assessment for existing substances, Directive 98/8/EC of the European Parliament and of the Council concerning the placing of biocidal products on the market Part II. EUR 20418 EN/2. Brussels, Belgium.

Hoenig JM, Heisey DM. 2001. The abuse of power: The pervasive fallacy of power calculations for data analysis. *Am Stat* 55:19–24.

Hothorn LA. 2014. Statistical evaluation of toxicological bioassays—A review. *Toxicol Res* 3:418–432.

Hothorn LA, Hasler M. 2008. Proof of hazard and proof of safety in toxicological studies using simultaneous confidence intervals for differences and ratios to control. *J Biopharm Stat* 18:915–933.

Landis WG, Chapman PM. 2011. Well past time to stop using NOELs and LOELs. *Integr Environ Assess Manag* 4:vi–viii.

McBride GB, Loftis JC, Adkins NC. 1993. What do significance tests really tell us about the environment? *Environ Manage* 4:423–432.

McDonald LL, Erickson WP. 1994. Testing for bioequivalence in field studies: Has a disturbed site been adequately reclaimed? In Fletcher DJ, Manly BFJ, eds, *Statistics in Ecology and Environmental Monitoring*. Otago Conference Series 2. University of Otago, Dunedin, New Zealand, pp 183–197.

Moore T, Canton S, Grimes M. 2000. Investigating the incidence of type-I errors for chronic whole effluent toxicity testing using *Ceriodaphnia dubia*. *Environ Toxicol Chem* 19:118–122.

New Mexico Energy, Minerals and Natural Resources Department, Mining and Minerals Division. 1999. Coal Mine Reclamation Program Vegetation Standards. Attachment 1. Santa Fe, NM, USA.

Organisation for Economic Co-Operation and Development. 2006. Current approaches in the statistical analysis of ecotoxicity data: A guidance to application. Series on Testing and Assessment, No. 54. ENV/JM/MONO(2006)18. Paris, France.

OSPAR Commission. 2007. Practical guidance document on whole effluent assessment. Publication 316/2007. London, UK.

Parkhurst DF. 2001. Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation. *BioScience* 51:1051–1057.

Power EA, Boumfrey RS. 2004. International trends in bioassay use for effluent management. *Ecotoxicology* 13:377–398.

R Development Core Team. 2016. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruxton G. 2006. The unequal variance *t*-test is an underused alternative to Student's *t*-test and the Mann-Whitney U test. *Behav Ecol* 17:688–690.

Scroggins R, van Aggelen G, Schroeder J. 2002. Monitoring sublethal toxicity in effluent under the metal mining program. *Water Quality Research Journal of Canada* 37:279–294.

Streiner DL. 2003. Unicorns do exist: A tutorial on "proving" the null hypothesis. *Can J Psychiatry* 48:756–761.

UK Environment Agency. 2005. Monitoring Certification Scheme (MCERTS): Performance standard for laboratories undertaking direct toxicity assessment of effluents. Ver 1. London, UK.

UK Environment Agency. 2007. The direct toxicity assessment of aqueous environmental samples using the juvenile *Daphnia magna* immobilization tests. Methods for the examination of waters and associated materials. London, UK.

US Environmental Protection Agency. 1995. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to west coast marine and estuarine organisms. EPA/600/R-95-136. Cincinnati, OH.

US Environmental Protection Agency. 2000. Understanding and accounting for method variability in whole effluent toxicity applications under the national pollutant discharge elimination system program. EPA/833/R-00/003. Washington, DC.

US Environmental Protection Agency. 2002a. Guidelines establishing test procedures for the analysis of pollutants; whole effluent toxicity test methods; final rule. *Fed Reg* 67:40 CFR 136.

US Environmental Protection Agency. 2002b. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to freshwater organisms. 3rd edition. EPA/821/R-02/013. Washington, DC.

US Environmental Protection Agency. 2010. National pollutant discharge elimination system test of significant toxicity technical document. EPA/833/R-10/004. Washington, DC.

US Environmental Protection Agency. 2016. Aquatic life ambient water quality criterion for selenium—Freshwater. EPA/822/R-16-006. Washington, DC.

Van der Vliet L, Taylor LN, Scroggins R. 2012. NOEC: Notable oversight of enlightened Canadians: A response to Van Dam et al. (2012). *Integr Environ Assess Manag* 8:397–398.

Warren-Hicks WJ, Parkhurst BR. 1996. Issues in whole effluent toxicity test uncertainty analysis. In Grothe DR, Dickson KL, Reed-Judkins DK, eds, *Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving System Impacts*. Special SETAC Publication. SETAC, Pensacola, FL, USA, pp 180–189.

Warren-Hicks WJ, Parkhurst BR, Moore DJ, Teed RS, Baird RB, Berger R, Denton DL, Pletl JJ. 2000. Assessment of whole effluent toxicity test variability: Partitioning sources of variability. *Environ Toxicol Chem* 19:94–104.

Zheng L, Diamond J, Denton DL. 2013. Evaluation of whole effluent toxicity data characteristics and use of Welch's *t*-test in the test of significant toxicity analysis. *Environ Toxicol Chem* 32:468–474.

<<ENOTE>>AQ1: Square brackets OK as added here: [100 × (25 − 20)]/25, or 20%.?

<<ENOTE>>AQ2: Please cross-ref to specific subheading add “in section xxx”.

<<ENOTE>>AQ3: “TAC” OK as spelled out as test acceptability criterion?

<<ENOTE>>AQ4: Figures 3 and 4 have been renumbered to reflect order of discussion. Manuscript cited/discussed Figures 1, 2, then 4 and 3. Please check all text citations.

<<ENOTE>>AQ5: OK as changed to “RWC” from “IWC”? Or spell out “IWC,” which was not defined or used elsewhere in the article.

<<ENOTE>>AQ6: CA State Water: Please add date accessed: [cited year month day].

[HC1]Typesetter: Please italicize n in n = x throughout figure 1 and put a space either side of the = sign.
[HC2]Typesetter: Please set n in italics throughout figure 2 in “n = x”. Also please change “Percent effect” to “Percentage effect”.
[HC3]Typesetter: Please change Percent effect to Percentage effect in Figure 4.

Table 1. Summary statistics of control data for the *Ceriodaphnia dubia* reproduction test endpoint for each laboratory in the study[a]

Laboratory	No. tests	Mean	SD	Percentiles of *CV*						
				0%	10%	25%	50%	75%	90%	100%
After 2012 (CIWQS)										
A	43	19.8	4.2	0.10	0.16	0.20	0.23	0.28	0.34	0.47
B	18	32.4	6.4	0.08	0.10	0.11	0.15	0.25	0.31	0.44
C	20	26.6	4.1	0.04	0.10	0.14	0.20	0.27	0.36	0.39
D (D2)	57	40.0	7.1	0.04	0.06	0.09	0.10	0.17	0.29	0.49
E	22	25.1	2.9	0.05	0.08	0.09	0.11	0.21	0.27	0.33
F	20	32.7	3.2	0.05	0.06	0.09	0.11	0.16	0.27	0.30
Before 2012 (test drive)										
D (D1)	30	31.6	5.2	0.04	0.09	0.13	0.17	0.31	0.37	0.51
G	17	29.8	6.0	0.05	0.06	0.09	0.09	0.13	0.16	0.52
H	17	26.2	6.8	0.04	0.07	0.09	0.10	0.16	0.34	0.57

[a] Laboratory D modified laboratory practices and quality controls during the time frame of the present study; D1 represents laboratory statistics prior to laboratory refinements, and D2 represents after laboratory refinements were instituted.

CV = coefficient of variation; CIWQS = California Integrated Water Quality System; SD = standard deviation.

Table 2. Comparison of percentiles for coefficients of variation of control *Ceriodaphnia* reproduction between the national study (US Environmental Protection Agency 2010) and the present study

Percentile	US Environmental Protection Agency (2010), TST technical document	Present study
0%	—	0.036
10%	0.08	0.076
25%	0.10	0.097
50%	0.15	0.147
75%	0.24	0.244
90%	0.35	0.332
100%	—	0.568
No. tests	792	244
No. laboratories	44	8

Table 3. Number of replicates required for test of significant toxicity to achieve a 5% probability of declaring toxicity at 10% effect[a]

Laboratory	Observed control *CV*			Number of replicates required	Probability (declaring toxic) at 10% effect with 10 replicates
	Average	75th percentile	Number of tests		
G	0.13	0.13	17	8	0.037
F	0.13	0.16	20	11	0.052
D2 [b]	0.15	0.17	57	13	0.071
E	0.15	0.21	22	20	0.070
H [c]	0.17	0.16	17	11	0.101
B	0.18	0.25	18	27	0.100
C	0.21	0.27	20	32	0.151
A	0.24	0.28	43	35	0.208

[a] The number of replicates is based on a mathematical power calculation using the laboratory's 75th percentile coefficient of variation (*CV)* and assumes that the standard deviation is the same for control and receiving water concentration. Probability of declaring toxicity with 10 replicates at 10% effect is based on resampling each laboratory's data; it is more closely related to average *CV*.

[b] Data for laboratory "D" from before 2012 ("D1" in Table 1) were excluded because laboratory practices and quality assurance/quality control were improved after 2012.

[c] The average (sample mean) of *CV* for laboratory H exceeds the sample 75th percentile, which is not improbable given $n = 17$ and some skewness (see Supplemental Data Part 3 for more detail).

Figure [HC1]1. Proportion of *Ceriodaphnia* reproduction tests that would be declared toxic using the test of significant toxicity as a function of the control coefficient of variation. Plots on the left show results of (nonparametric) resampling of 244 individual tests' control replicates as dots, superimposed on the (parametric) mathematical calculation shown as a line, to establish how closely these 2 agree. Each dot is an average of 10 000 resamples from one of the 244 controls. Plots on the right show mathematical power calculations for n = 10, 20, and 30 replicates.

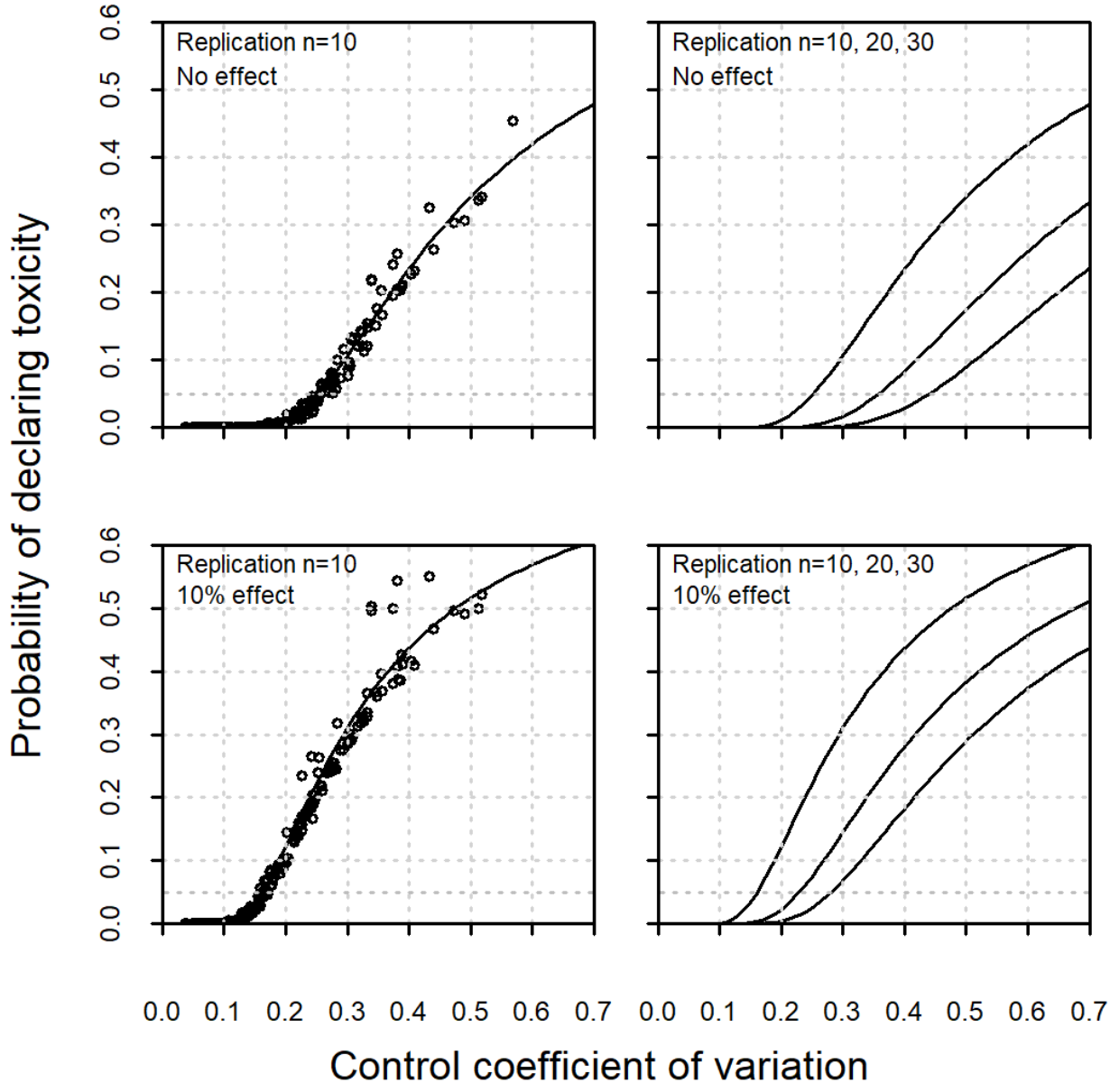

Replication n=10
No effect
Replication n=10, 20, 30
No effect
Replication n=10
10% effect
Replication n=10, 20, 30
10% effect
Probability of declaring toxicity
0.0 0.1 0.2 0.3 0.4 0.5 0.6
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
Control coefficient of variation

Figure [HC2]2. Probability of declaring a sample toxic using the no-observed-effect concentration (NOEC) and test of significant toxicity (TST) based on simulating 10 000 whole-effluent toxicity tests at each of various percentage effect parameter values (horizontal axis), 4 values of control coefficient of variation parameter, and 3 values for number of test replicates. Gray horizontal line shows probability of 0.05. Solid curves represent TST and broken curves, NOEC. *CV* = coefficient of variation.

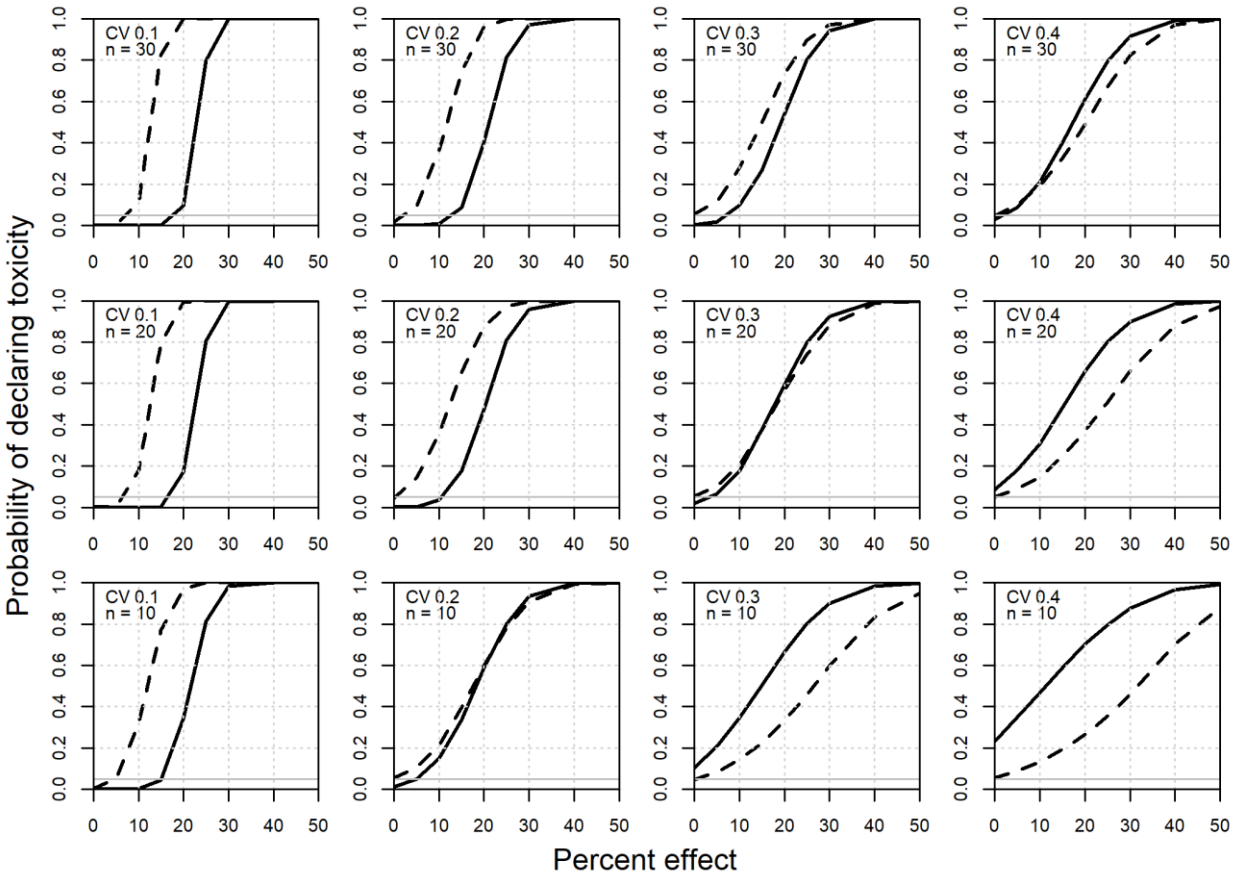

Probability of declaring toxicity
Percent effect
CV 0.1
n = 30
CV 0.2
n = 30
CV 0.3
n = 30
CV 0.4
n = 30
CV 0.1
n = 20
CV 0.2
n = 20
CV 0.3
n = 20
CV 0.4
n = 20
CV 0.1
n = 10
CV 0.2
n = 10
CV 0.3
n = 10
CV 0.4
n = 10
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50

Figure 3. Probability of declaring a sample toxic using the test of significant toxicity with the *Ceriodaphnia* reproduction test, in relation to number of replicates and control coefficient of variation (*CV)*. Four curves show *CV* = 0.1 (lowest curve, solid line), 0.2, 0.3, and 0.4 (highest curve, dotted line). Percentage effect parameter is zero in the upper plot and 10% in the lower plot.

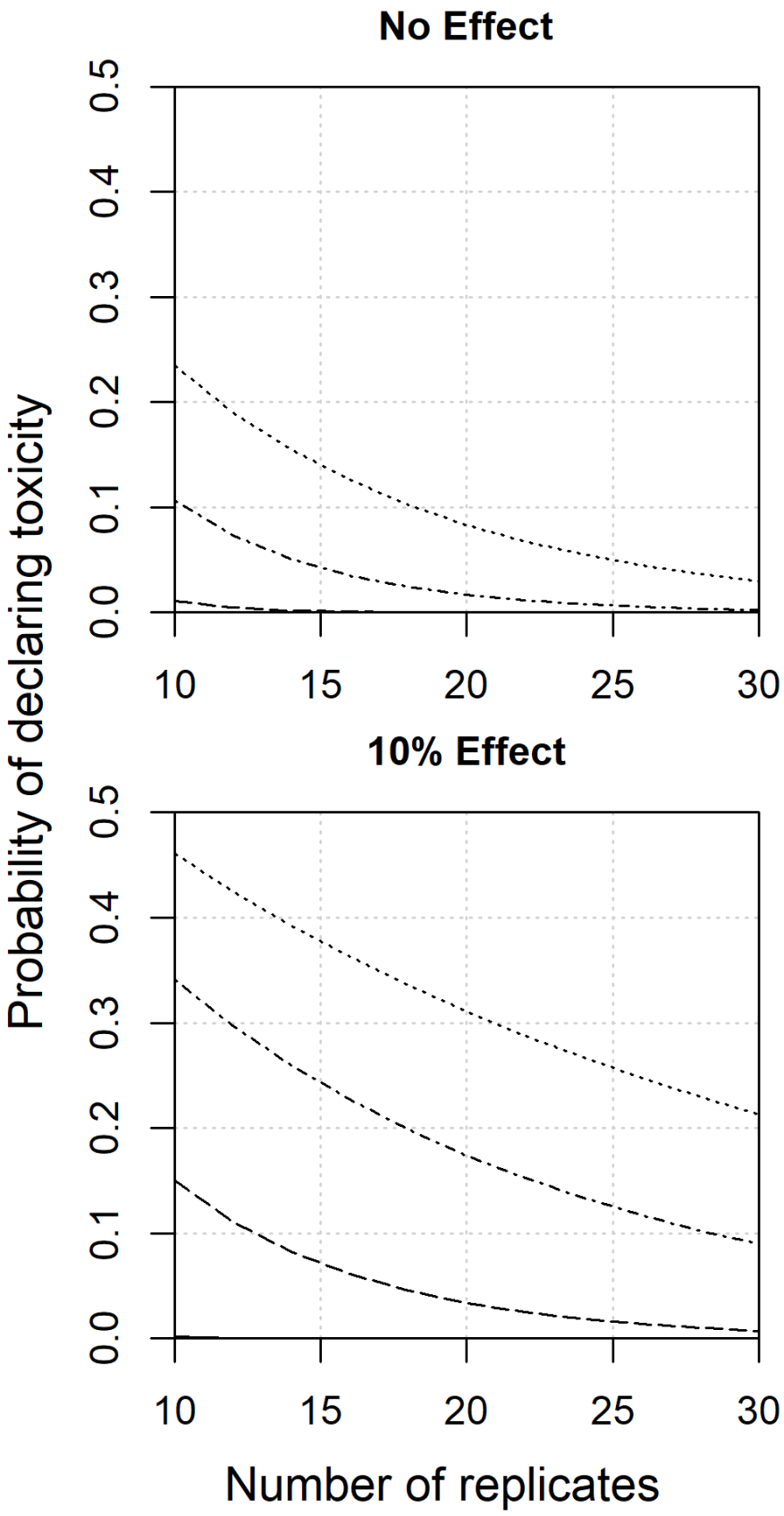
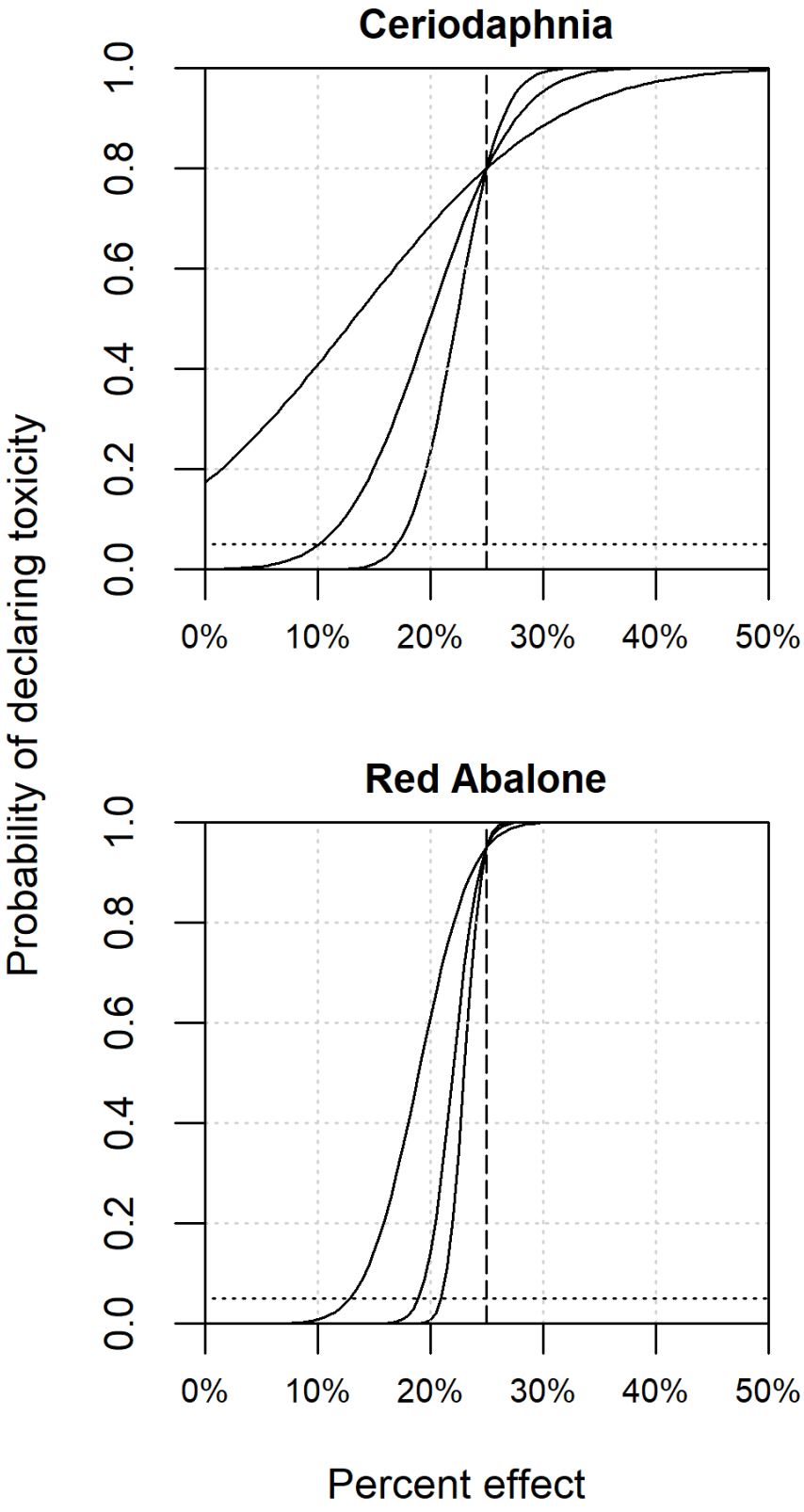

No Effect
10% Effect
Probability of declaring toxicity
0.0
0.1
0.2
0.3
0.4
0.5
10
15
20
25
30
Number of replicates

Figure [HC3]4. Probability of declaring a sample toxic for *Ceriodaphnia* and red abalone with the test of significant toxicity (using the minimum number of replicates specified for each whole-effluent toxicity test). Vertical dashed line shows 25% effect. Steeper (less divergent) curves have smaller coefficients of variation (*CV)*. The 3 *CV* values correspond to low, medium, and high *CV*s from a national sample (specifically, the 10th, 50th, and 90th percentiles of *CV*s achieved for each toxicity test method; US Environmental Protection Agency 2010). These percentiles are 0.08, 0.15, and 0.35 for *Ceriodaphnia* and 0.02, 0.03, and 0.06 for red abalone, respectively.


Ceriodaphnia
Red Abalone
Probability of declaring toxicity
0.0
0.2
0.4
0.6
0.8
1.0
0%
10%
20%
30%
40%
50%
Percent effect