Environmental Toxicology and Chemistry—Volume 39, Number 11—pp. 2109–2123, 2020
Received: 20 March 2020 | Revised: 11 May 2020 | Accepted: 6 August 2020

Critical Review

The Minimum Detectable Difference (MDD) Concept for Establishing Trust in Nonsignificant Results: A Critical Review

Magdalena M. Mair,[a,*] Mira Kattwinkel,[b] Oliver Jakoby,[c] and Florian Hartig[a]

[a]Faculty of Biology and Pre-Clinical Medicine, Theoretical Ecology, University of Regensburg, Regensburg, Germany
[b]Institute for Environmental Sciences (iES), University of Koblenz-Landau, Landau, Germany
[c]RIFCON GmbH, Hirschberg, Germany

Abstract: Current regulatory guidelines for pesticide risk assessment recommend that nonsignificant results should be complemented by the minimum detectable difference (MDD), a statistical indicator that is used to decide whether the experiment could have detected biologically relevant effects. We review the statistical theory of the MDD and perform simulations to understand its properties and error rates. Most importantly, we compare the skill of the MDD in distinguishing between true and false negatives (i.e., type II errors) with 2 alternatives: the minimum detectable effect (MDE), an indicator based on a post hoc power analysis common in medical studies; and confidence intervals (CIs). Our results demonstrate that MDD and MDE only differ in that the power of the MDD depends on the sample size. Moreover, although both MDD and MDE have some skill in distinguishing between false negatives and true absence of an effect, they do not perform as well as using CI upper bounds to establish trust in a nonsignificant result. The reason is that, unlike the CI, neither MDD nor MDE consider the estimated effect size in their calculation. We also show that MDD and MDE are no better than CIs in identifying larger effects among the false negatives. We conclude that, although MDDs are useful, CIs are preferable for deciding whether to treat a nonsignificant test result as a true negative, or for determining an upper bound for an unknown true effect. *Environ Toxicol Chem* 2020;39:2109–2123. © 2020 The Authors. *Environmental Toxicology and Chemistry* published by Wiley Periodicals LLC on behalf of SETAC.

Keywords: Minimum significant difference; Least significant difference; Minimum detectable change; Post hoc power; Statistical ecotoxicology; Risk assessment

INTRODUCTION

Conventional agriculture relies heavily on the use of pesticides (Tilman et al. 2002; European Commission 2020; Food and Agriculture Organization of the United Nations 2020). To limit the negative side effects of these pesticides, risk assessment examines the effect of pesticides on nontarget organisms and compares these effects with the predicted pesticide exposure in the field. One of the most common analyses in aquatic risk assessment is the calculation of regulatory acceptable concentrations (European Food Safety Authority 2013; European Food Safety Authority PPR Panel 2013). The regulatory acceptable concentrations are pesticide concentrations that impose only negligible (European Food Safety Authority PPR Panel 2013) or acceptable (SANCO 2002a, 2002b; European Commission 2009) effects on nontarget organisms (ecological threshold option-regulatory acceptable concentration), or allow populations to recover within an acceptable (European Food Safety Authority PPR Panel 2013) or an ecologically relevant (SANCO 2002a) time period (ecological recovery option-regulatory acceptable concentration).

The definition of negligible in this context is usually that the toxic effect of a pesticide concentration is statistically nonsignificant compared to a control (see illustration in Figure 1). In practice, the most common approach is testing a range of pesticide concentrations against a pesticide-free control group. The regulatory acceptable concentration is then calculated based on the transition from the pesticide concentration showing a nonsignificant effect (no-observed-effect concentration [NOEC]) to the next higher concentration showing a significant effect (lowest-observed-effect concentration [LOEC]; note that this has been criticized repeatedly over the years: Laskowski 1995; Organisation for Economic

This article includes online-only Supplemental Data.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
* Address correspondence to magdalena.mair@ur.de
Published online 12 August 2020 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/etc.4847

© 2020 The Authors

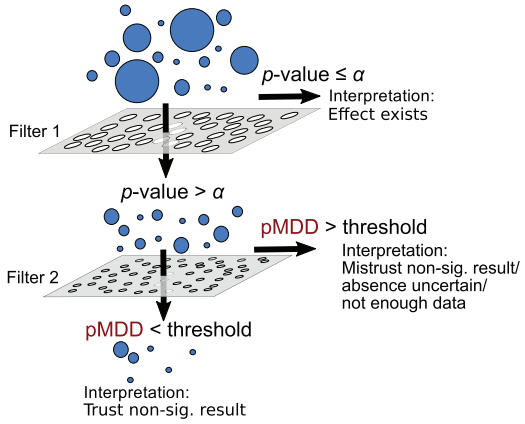

FIGURE 1: Illustration of the 2 filters that each experimental endpoint needs to pass to be considered showing "no effect" in the risk assessment. The first filter is a significance test, which is passed when $p > \alpha$ (not significant). The second filter requires a sufficiently low proportional minimum detectable difference (pMDD) and has the purpose of deciding whether the power of the experiment was sufficient to trust nonsignificant results that pass filter 1. Each circle represents one test (control vs treatment for one endpoint), and the circle sizes indicate different real effect sizes. Our illustration suggests that the filters exclude larger effect sizes, while letting smaller or null effects pass, which is what one would hope for. Later in our study, we used simulations to investigate how well the pMDD filter works for this purpose and compared it with the proportional upper bound of the confidence interval (pCI) and proportional minimum detectable effect with 80% power (pMDE) filters.

Co-operation and Development 1998; van Dam et al. 2012; Fox and Landis 2016).

The main concern regarding the use of significance tests for determining regulatory acceptable concentrations is the issue of power: the rate at which a significance test will detect a true effect generally depends on the power of the statistical design, which can be low to 0, for example, if the sample size is too small. This issue is particularly worrisome for higher tier experiments such as mesocosm studies, because these costly experiments are often conducted with low sample sizes (Touart 1994; Sanderson 2002; de Jong et al. 2006; Organisation for Economic Co-operation and Development 2007; Cabrera et al. 2016; Lemoine et al. 2016) and show large natural variation (Liber et al. 1992; Kraufvelin 1998). Both factors contribute to low power, and thus high type II error rates (i.e., false negatives; Organisation for Economic Co-operation and Development 2006; European Food Safety Authority PPR Panel 2013; Duquesne et al. 2020). To establish a reasonable regulatory framework, it is therefore essential to complement null hypothesis significance tests with minimum requirements on the power of the experiment (SANCO 2002a; Sanderson and Petersen 2002; de Jong et al. 2006; Organisation for Economic Co-operation and Development 2006; European Food Safety Authority 2013; European Food Safety Authority PPR Panel 2015), or find some other means to identify false negatives arising from low power.

Ensuring sufficient power can be addressed at 2 stages of an experiment: a priori power calculations are performed before the experiment, typically to determine the sample size necessary to detect a certain effect (O'Keefe 2007; Johnson et al. 2015). The second option is to calculate the so-called post hoc power, which means that power is determined after the experiment based on the collected data. Using post hoc power calculations to interpret nonsignificant results has been frequently recommended over the years (Fagley 1985; Cohen 1988; Steidl et al. 1997; Onwuegbuzie and Leech 2004; European Food Safety Authority PPR Panel 2015), but equally often criticized (Goodman and Berlin 1994; Zumbo and Hubley 1998; Heisey 2001; Lenth 2001; Colegrave and Ruxton 2003; Baguley 2004; O'Keefe 2007; Greenland 2012).

The criticism against post hoc power calculations depends on how those calculations are used. Their most naïve use is to determine the power of the experiment post hoc based on the estimated effect size and variance. However, statistical calculations show that if the p value is nonsignificant, a post hoc power calculation for the estimated effect size will always result in low power and thus does not add extra information on top of the p value (Hoenig and Heisey 2001; Baguley 2004; O'Keefe 2007; Greenland 2012).

The second, more common variant of post hoc power calculations is to use the estimated variance in the data for calculating the effect size that could have been detected with a specified power (typically in 80% of experiments). The resulting value, which is known as the minimum detectable effect (MDE), is more informative, but it also has shortcomings for the post hoc interpretation of nonsignificant results. Most importantly, the MDE is independent of the estimated effect size. As such, it informs about the theoretical power of the experiment to detect a certain effect, but the common practice of using the MDE value, or a threshold on the MDE, to decide whether to accept the nonsignificant result of the same experiment as a true negative can lead to counterintuitive behavior.

To understand this, consider, for example, 2 nonsignificant experiments with identical variance, but different estimated effect sizes. Because their variance is identical, both experiments would be assigned the same MDE (Figure 2), and thus the same level of trust in a scheme that determines whether to accept a nonsignificant result as a true negative based on the MDE. This result goes contrary to the intuition that the experiment with higher estimated effect size would provide more evidence for a true effect. Another example is the case in which both experiments estimate the same effect, but the second has a smaller variance and thus seems to provide more evidence for an, alas, small effect. Determining trust based on the MDE value, however, would lead to the opposite conclusion, because a smaller variance automatically increases power and thus leads to a smaller MDE. These counterintuitive examples of post hoc power analysis have been summarized earlier under the term "power approach paradox" (Hoenig and Heisey 2001).

For pesticide risk assessment, yet another indicator has been introduced to deal with the interpretation of power and nonsignificant results. The minimum detectable difference (MDD; sometimes also referred to as the minimum significant difference (MSD) (van der Hoeven 2008) is widely used in Europe and North America (Environment Canada 2005;

© 2020 The Authors

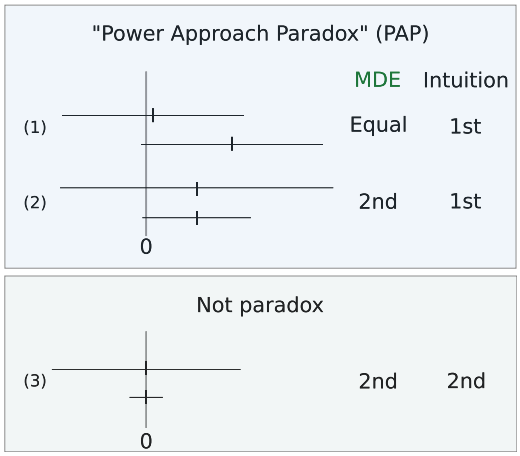

FIGURE 2: The power approach paradox (PAP) describes pairs of nonsignificant experiments for which the interpretation of post hoc power analysis contrasts statistical intuition. 1) For 2 experiments with equal variance (represented by horizontal lines) and different estimated effect sizes (vertical lines), the minimum detectable effect (MDE) considers both as equally indicative of a 0 effect, whereas intuition considers the experiment with lower estimated effect as a stronger indication for the absence of an effect. 2) For 2 experiments with equal estimated effect size but different variances, the interpretation of the MDE can also contradict intuition, 3) but this is not always the case (e.g., when the estimated effect is close to 0). 1st, 2nd, equal = which of the 2 experiments is interpreted as stronger evidence for a 0 effect.

European Food Safety Authority PPR Panel 2013; Brock et al. 2015), and its use has recently expanded from aquatic (European Food Safety Authority PPR Panel 2013; Brock et al. 2015; Nys et al. 2019) to terrestrial ecotoxicology (Scholz-Starke et al. 2013; Höss et al. 2014; Andrade et al. 2017; European Food Safety Authority PPR Panel 2017; Candolfi et al. 2018), as well as from simple statistical tests (van der Hoeven 2008; Brock et al. 2015) to more complex statistical models (Peters et al. 2016; Rolke et al. 2016).

The MDD is defined as the effect size that would have been just significant considering the sample size and variance measured in a conducted experiment (van der Hoeven 2008; Brock et al. 2015). In other words, the MDD is the effect size that corresponds to the critical value of the test statistic, which marks the border between significance and nonsignificance (see illustration in Figure 3). The MDD is principally measured in units of the observed response variable (endpoint; e.g., number of individuals), but to make it more comparable among different variables (e.g., different species in a community), it is then usually compared with the control mean. The latter yields the proportional MDD (pMDD), the percentage of reduction in species or taxon abundance (or any other endpoint) that would have just been significant. Lower pMDD values are frequently interpreted as indicating higher power of the experiment and, consequently, higher reliability of a nonsignificant test result.

Current regulatory guidelines for higher tier mesocosm and semifield studies recommend first performing null hypothesis significance tests to show that there is no evidence for adverse effects on nontarget organisms, followed by calculation of the pMDD (de Jong et al. 2008; European Food Safety Authority PPR Panel 2013; Brock et al. 2015; European Food Safety Authority 2019). To ensure sufficient study quality, regulatory guidelines can then specify minimum values for the pMDD (European Food Safety Authority PPR Panel 2013; Brock et al. 2015; Nys et al. 2019), which effectively constitutes a second filter (see Figure 1) that a nonsignificant result has to pass: if the pMDD is lower than the threshold, a nonsignificant test result is deemed to indicate a true absence of an effect at this specific pesticide concentration and time point (see description of effect classes in European Food Safety Authority PPR Panel 2013; Brock et al. 2015). In mesocosm studies for pesticide risk assessment in the European Union, endpoints are then assigned an effect class ranging from "not enough data" (class 0) to "no treatment-related effects" (class 1) to "pronounced long-term effects without recovery" (class 5B; Brock et al. 2000, 2015; de Jong et al. 2008; European Food Safety Authority PPR Panel 2013). Effect class assignment is based on the pMDDs found at different concentrations and time points and in addition takes into account the plausibility of effects at these measurement points (e.g., following a consistent dose–response relationship, timing of effects after pesticide application, etc.). In this context, *p* values are thus used to determine concentrations and time points with effects (for the determination of the LOEC), and pMDD values are suggested to be used post hoc as indicators for the level of evidence regarding the absence of real effects in subsequent measurements with nonsignificant *p* values (NOEC; see second filter in Figure 1).

Although the definition of the MDD is mathematically clear, its appropriateness as a "true negative filter" or a "level of evidence indicator" raises many of the same questions that have been asked previously about post hoc power analysis. In particular, although significance tests clearly act as a filter on real effects (because *p* values are correlated to real effect size), it is unclear whether the same holds true for the MDD, which, similar to the MDE, is independent of the estimated effect. Moreover, it is noteworthy that the MDD is used almost exclusively in ecotoxicology, whereas medical studies and other fields usually use MDEs or confidence intervals (CIs) to interpret nonsignificant results. The CIs are measures of uncertainty, which are usually reported in addition to effect sizes to indicate the precision of these estimates. Similar to the calculation of the *p* value, the calculation of the CI takes into account the estimated effect size. Because the use of the MDD in ecotoxicological research has recently expanded and partly substitutes for the more widely used MDE and CI, it seems useful to examine the relationship of the MDD to these statistical indicators more closely.

In the present study, we review the theory of the MDD, describe its relationship to other statistical indicators (*p* value, MDE, and CI) and statistical properties (power), and subsequently perform simulations to examine the performance of the MDD in a context typical for higher tier aquatic ecotoxicological studies. Specifically, we used simulations to compare the performance of proportional (p)CIs (i.e., the CI upper bounds related to control mean), proportional (p)MDEs

 © 2020 The Authors

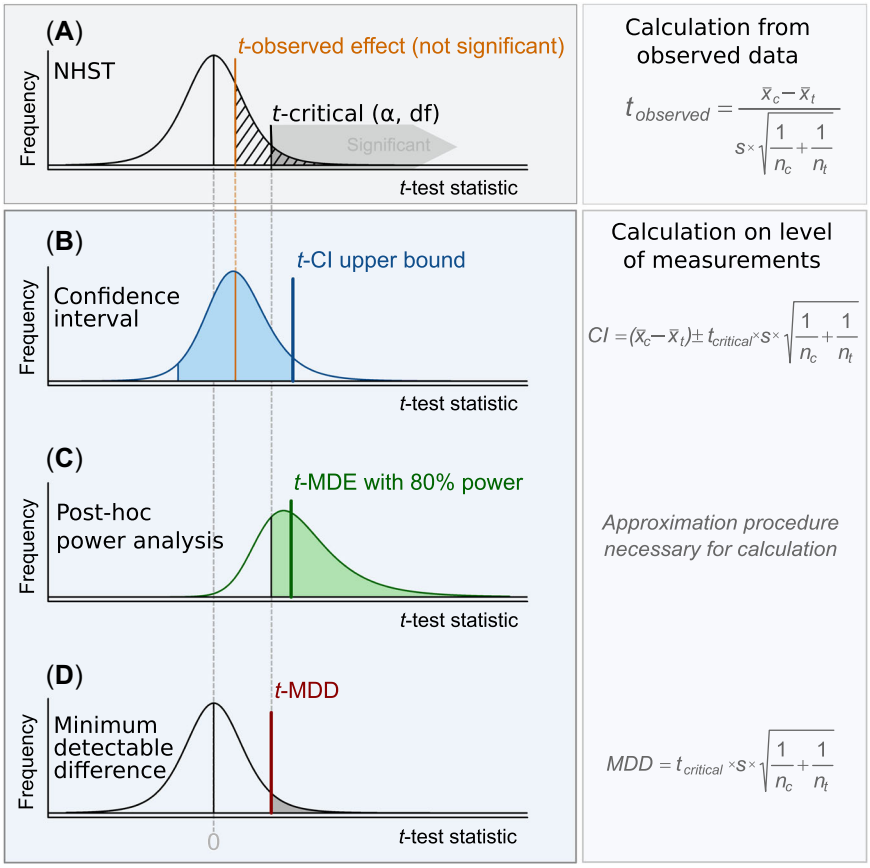

FIGURE 3: An illustration of the statistical concepts discussed, using the *t* test as an example. (**A**) The distribution of the test statistic under the null hypothesis H_0 is used to calculate the *p* value (the probability of obtaining a *t* larger or equal to *t*-observed under H_0; hatched area). The test becomes significant if *p* is smaller than the significance level *α*. (**B**) The confidence interval is based on calculating the *t*-distribution around the estimated effect. (**C**) The 80% minimum detectable effect (MDE) is obtained by requiring that 80% of *t* values are larger than *t*-critical (shaded area) under the assumption that the MDE is the true effect (the latter results in a noncentral *t*-distribution). (**D**) The minimum detectable difference (MDD) corresponds simply to *t*-critical. The right column shows the formulae used to obtain the respective indicators. x_c, x_t = control and treatment group means; n_c, n_t = sample sizes of control and treatment groups; *s* = residual standard deviation (square root of residual variance); *t*-critical = *t* value that marks the border between significant and nonsignificant *t* values (it depends on the *α*-level and degrees of freedom *df*). NHST = null hypothesis significance test.

(i.e., MDEs from post hoc power analysis related to control mean) with 80% power, and proportional (p)MDDs (i.e., MDDs related to control mean) in 2 different tasks: 1) their ability to discriminate between true absence of effects and false negatives (type II errors); and 2) their sensitivity to real effect size when applied as secondary filters on experiments with nonsignificant results.

A SHORT SUMMARY OF THE THEORY OF *p* VALUES, MDD, MDE, AND CI

Null hypothesis significance tests

Despite a decade-long discussion about their use and misuse (e.g., Newman 2008; Greenland et al. 2016; Erickson and Rattner 2020), null hypothesis significance tests are still the most widely used statistical method for determining the existence of an effect and are required in most regulatory guidelines.

A null hypothesis significance test first requires the definition of a null hypothesis H_0 (typically no effect). The next step is to define a test statistic, which summarizes certain desired aspects of the data, typically related to the effect size. Each value of the test statistic can thus be assigned a corresponding effect size. The expected distribution of the test statistic under H_0 (see black curve in Figure 3A) is then calculated, which provides an idea of the values of the test statistic that would be likely to occur, if H_0 (typically: no effect) was true. Based on the distribution of the test statistic under H_0, we can now calculate the *p* value (hatched area below the frequency distribution of *t* in Figure 3A), defined as the probability of obtaining the observed value of the test statistic or extremer under H_0. Thus, the *p* value is a measure of how strongly the observed data deviates from the values expected under H_0, where the deviation is quantified by the test statistic.

The procedure of null hypothesis significance tests now prescribes that, for *p* values lower than a chosen significance level (*α*, usually 5%), the null hypothesis is rejected, and the effect is considered significant. This is true for all values beyond the critical value where *p* equals *α* (gray shading in

© 2020 The Authors

Figure 3A). If the *p* value is larger than α, the null hypothesis cannot be rejected, and the estimated effect is considered nonsignificant.

Type II error rates in null hypothesis significance tests are not controlled

A well-known implication of the null hypothesis significance test procedure, and critical for ecotoxicological risk assessment, is that null hypothesis significance tests control only the rate of false positives (type I errors). The very definition of the *p* value guarantees that the type I error rate is fixed at a level of α (typically 5%). The type II error rate (the rate of false negatives), or its counterpart, power (the rate at which true effects are detected), however, is uncontrolled and depends in general on sample size, true effect size, and variance in the data.

Due to the asymmetry between the control of type I and type II errors, a significant result constitutes relatively clear evidence in favor of an effect, whereas a nonsignificant result can as easily be explained by low power as by the true absence of an effect. In other words, null hypothesis significance tests are designed primarily to show effects, not their absence. If the latter is the purpose of the experiment, as is the case in ecotoxicological risk assessment, we have to complement them by some other statistical procedure to control type II error rates.

Approaches to control type II error rates post hoc

Because the lacking control of power is the main problem for examining the absence of effects via null hypothesis significance tests, it seems natural to complement them by post hoc power calculations. In practice, the following 2 indicators are regularly used for this purpose: the MDE and (in ecotoxicological risk assessment) the MDD. An alternative route is to simply consider the CI, which provides information about the precision of the estimate, and thus about the upper bound for possible real effects. We will now introduce all 3 indicators (Figure 3B–D).

MDE. In the context of null hypothesis significance tests, the MDE is defined as the effect size that can be detected with specified power in a given test (Figure 3C). When calculating the MDE post hoc, the observed variance is usually used as an estimate of the true variance in the experiment, and in practice, a power of 80% is usually targeted (European Food Safety Authority 2013; European Food Safety Authority PPR Panel 2015). A lower MDE is interpreted as an indication for higher power, but note that because the calculation of the MDE neglects the estimated effect size, its direct interpretation as evidence for the absence of an effect is problematic (Figure 2; see also Hoenig and Heisey 2001). This behavior has been termed a paradox, but note that it is not paradoxical when considering that the MDE informs generally about the power of the experiment. Due to its neglect of the estimated effect size, however, the MDE is only imperfectly correlated with the probability that an effect is present in the given situation, and should not be interpreted as such.

MDD. Similar to the MDE, the MDD is a measure of power. It is defined as the effect size that would have been just significant in the conducted experiment (Figure 3D). The MDD is calculated by taking the critical value of the test statistic (here: *t*-critical = *t*-MDD), which marks the border between nonsignificance and significance, and calculate the corresponding effect size (in units of the measured endpoint) by rearranging the equation used for the calculation of the test statistic (see equation in Figure 3A). For the *t* test, the MDD is therefore calculated based on the variances in the treatment groups, the degrees of freedom (i.e., sample size), and the selected α-level (see equation in Figure 2D; Brock et al. 2015) and equals the upper bound of the CI (see the next section) minus the estimated effect (compare equations in Figure 3B,D). Thus, for the *t* test, one can imagine the MDD as being similar to a CI, just constructed around H_0, typically no effect. Identically to the MDE, the MDD does not consider the estimated effect size in its mathematical definition and should thus not directly be interpreted as indicating the probability of a 0 effect (Figure 2; see also Hoenig and Heisey 2001).

CI. Unlike MDD and MDE, CIs are not primarily intended to measure power, but rather to assert the precision of an estimated effect (Figure 3B). When conducting the same experiment repeatedly, the CI will include the real effect size at a fixed rate that equals the chosen confidence level (typically 95%). For example, by calculating a 2-sided CI with a confidence level of 90%, the real effect will be larger than the upper CI bound and smaller than the lower CI bound in 5% of the experiments each (Figure 3B). For the *t* test, the range of the CI is calculated based on the critical *t*-value, sample size, measured variance, and estimated effect size (see equation in Figure 3B). For nonsignificant results, the upper bound of the CI can be interpreted as an upper bound for a possible effect, with smaller values increasing trust in nonsignificant results. In contrast to the MDE, among 2 nonsignificant experiments with equal variance, the CI will assign more trust in a 0 effect to the experiment with lower estimated effect (compare with first case in Figure 2).

Regardless of the somewhat more intuitive behavior of the CI, it is important to understand that neither CI, nor MDE, or MDD are designed to measure the posterior probability of the null hypothesis. The MDE and MDD measure power, and the CI measures the precision of the estimated effect size. Other statistical indicators exist that aim at directly estimating the probability of the null hypothesis, and we address those in the *Discussion* section. Using these indicators, however, would require abandoning the use of standard hypothesis tests. There are reasons for doing so, but there are also reasons to avoid this, including the wide acceptance and familiarity of the field with the conventional null hypothesis significance testing procedure. When one decides to remain within the standard null

 © 2020 The Authors

hypothesis significance testing framework, which is what we assume in the present review, what CI, MDE, or MDD deliver is not the probability of H_0, but rather a threshold to control type II error rates. Thus, they must be assessed in their ability to identify type II errors, and not in their ability to determine the probability of a 0 effect. In that sense, we find the power approach paradox (see also Figure 2) important, but not a categorical counterargument against the use of any of the 3 indicators.

CLARIFYING A COMMON MISUNDERSTANDING: THE MDD DOES NOT HAVE CONTROLLED POWER

After this broad exposition about approaches to control type II errors in the null hypothesis significance testing framework, we want to direct our attention again more closely to the MDD, which, after all, is the main approach recommended to avoid type II errors in ecotoxicological risk assessment. Our principal question will be whether the MDD is indeed suited for this goal. Before addressing that, however, we want to clarify some misunderstandings regarding its interpretation, which also relates to the distinction between the MDD and the MDE.

Given the definition of the MDD as the effect size that would have been significant, it is tempting to think that 1) there would be a high power to detect an effect of the size of the MDD (as suggested, e.g., in Andrade et al. 2017; Green et al. 2018) and 2) the power to detect an effect of the size of the MDD is fixed (as suggested, e.g., in Duquesne et al. 2020). Neither is true.

In fact, the MDD has relatively low power. To see that, assume that the real effect is equal to the calculated MDD. For a one-sided *t* test, the corresponding power is the area below the *t*-distribution from *t*-critical (which corresponds to the MDD; see Figure 3) to infinity (red area below the solid line in Figure 4A). If the *t*-distribution for a real effect of size MDD (i.e., around *t*-critical) was symmetric (which is true asymptotically), the corresponding power would be 50%, because the values of the *t*-distribution would fall in equal proportions on either side of *t*-critical (Hoenig and Heisey 2001; Duquesne et al. 2020). Thus, asymptotically, the MDD has a power of 50%, which would not be considered high according to normal statistical standards.

When one moves beyond the asymptotic argument (i.e., considering small sample sizes, which is the more relevant case for ecotoxicological risk assessments), the situation is slightly more complicated. For positive real effects, the *t*-distribution becomes increasingly skewed as the sample size decreases (noncentral *t*-distribution; Johnson and Welch 1940; Owen 1965; see the positive skew in the *t*-distribution around *t*-MDD in Figure 4A). A positive skew means that the distribution is flatter on the right than on the left side and, consequently, the area below the *t*-distribution is larger to the right than to the left side of *t*-critical. Vice versa, for negative real effects, the *t*-distribution becomes increasingly

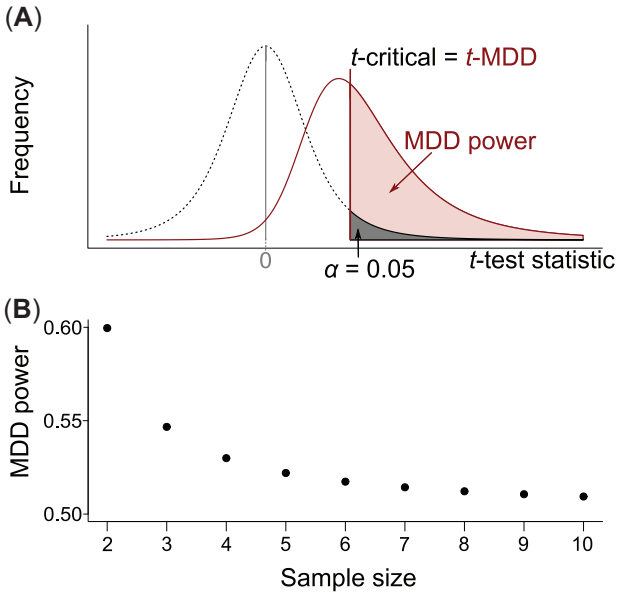

FIGURE 4: (**A**) Illustration of the distribution of the *t* statistic under the assumption that the true effect size is equal to the minimum detectable difference (MDD). The black dotted line is the central *t*-distribution that is used to calculate the *p* value, which will be significant, if *t*-observed is larger than *t*-critical. The red line depicts the noncentral *t* distribution around *t*-critical (real effect of size MDD). The red shaded area below the noncentral *t*-distribution to the right of *t*-critical equals the power of the MDD. (**B**) With increasing sample size, the *t*-distribution becomes more symmetric and thus MDD power approaches 50%.

negatively skewed for smaller sample sizes and thus flatter on the left side.

As a result of this skew, the power to detect an effect of size MDD depends on the sample size, because the latter controls the skew of the *t*-distribution around *t*-critical. In general, the power of detecting an effect of the size of the MDD increases with decreasing sample size, and practical values in ecotoxicological risk assessment probably range between 50 and 60% (Figure 4B; R code is provided in the Supplemental Data 1.1). Moreover, besides the sample size, the skew of the *t*-distribution also depends on whether variances of control and treatment are assumed to be equal, or nonequal (see Figure S1, Supplemental Data 1.2).

CAN THE MDD DISCRIMINATE BETWEEN TRUE AND FALSE NEGATIVE TESTS?

The relatively low power of the MDD may be slightly disconcerting, but this alone does not constitute a problem for using the MDD as a measure of trust in nonsignificant results. As long as regulators keep in mind that the power of the MDD is relatively low and set the minimum requirements on the MDD accordingly, the MDD filter can in principle be made as sharp as desired (see illustration in Figure 1). The important question is whether the MDD filter can distinguish between false negatives and true absence of effects in the experiments it is presented with. What we therefore have to investigate is the

© 2020 The Authors

skill of the MDD to discriminate between true zero effects and false negatives (type II errors), relative to alternative statistical measures.

Simulation of experiments

To examine this question, we simulated a large number of experiments typical for MDD applications (e.g., in mesocosm experiments) and evaluated the error rates of pCI (proportional CI upper bound), pMDE (proportional MDE with 80% power), and pMDD for deciding whether to trust a nonsignificant test result, that is accept it as a zero effect. The response variable in the experiments was species abundances (i.e., count data), and we assumed that these data would be log-transformed to make them approximately normal, which is a common procedure in practice. Moreover, for simplicity, we limited our design to one control and one treatment group, and only one time point and species. Each experiment can thus be analyzed by a single *t* test. We note that there are more appropriate ways to deal with the analysis of count data (see, for example, Szöcs and Schäfer 2015; Lehmann et al. 2016), but we chose the commonly used log-transformation because it allows us to continue using the *t* test and because this procedure is frequently used in this context (Brock et al. 2015). The results can be expected to generalize to situations in which several pesticide concentrations are tested against the control, for example, via a Williams or Dunnett's test, which control for multiple testing (Williams 1972; Kennedy et al. 1999; de Jong et al. 2006; European Food Safety Authority PPR Panel 2013; Brock et al. 2015), as well as via other statistical tests.

We simulated data by drawing random samples from a negative binomial distribution, which creates the type of overdispersed count data that is typical for such experiments. In a first batch, we simulated 2000 experiments, half of which had a real effect size of 0 (i.e., no real effect/no difference in mean abundance between treatment and control). For the other half, the real effect size was set to 0.5 (i.e., 50% reduction in mean abundance in the treatment group compared with the control). Sample size, real species abundance in the control (i.e., distribution mean), and the dispersion parameter of the negative binomial distribution (theta) were drawn from uniform distributions to achieve generality of the results. The sample size ranged from 3 to 10, the dispersion parameter theta ranged from 1 to 50, and the real mean abundance in the control group ranged from 10 to 100 (only integers). The real mean abundance in the treatment group was then either identical (no effect) or 50% reduced (effect size of 0.5).

To test whether simulation outcomes were influenced by real effect size, we conducted additional simulations in which we compared zero effects with effects of 0.2 (20% reduction in treatment group) and zero effects with effects of 0.8 (80% reduction in treatment group).

For each experiment, the data were $\ln(2 \times x + 1)$-transformed (following Brock et al. 2015), and a one-sided *t* test (control mean larger than treatment mean) with $\alpha = 0.05$ was performed, followed by calculation of the upper bound of a one-sided 95% CI (for results using 75% CIs see the Supplemental Data 3.4), MDE with 80% power, and MDD (an R function for the calculation of the MDD is provided in the Supplemental Data 2). All 3 statistical indicators were then back-transformed, related to the back-transformed control mean, and expressed as percentage of change in species abundance (pCI, pMDE, and pMDD; see Brock et al. 2015; note that calculating the mean of the transformed data followed by back-transformation is not equal to calculating the mean of the untransformed observed data, unless the variance in the transformed data is 0: see Jensen 1906; Green et al. 2018).

We then filtered all experiments first based on the *p* value. Experiments with $p \leq 0.05$ were interpreted as "effect exists" and not considered further. For results that were not significant ($p > 0.05$), we applied all 3 statistical indicators (pCI/pMDE/pMDD) with different threshold levels ranging from 30 to 100 (see illustration in Figure 1 and schematic depiction of the experimental approach in Figure 5A). Nonsignificant experiments with pCI/pMDE/pMDD larger than the threshold level were interpreted as "mistrust nonsignificant result/absence of effect uncertain/not enough data," whereas nonsignificant experiments with pCI/pMDE/pMDD smaller than the threshold level were interpreted as "trust nonsignificant result."

Based on these interpretations, error rates were calculated for all indicators and threshold levels. To avoid confusion with type I and type II errors, which correspond to the error rates of *p* values, we use the term "false trust rate" to refer to the rate at which an indicator suggested to trust a nonsignificant result despite the presence of a real effect, and "false mistrust rate" to refer to the rate at which an indicator erroneously suggested to mistrust a nonsignificant result (Figure 5A; R code is provided in the Supplemental Data 3.1 and 3.2).

Because there is no clear correspondence of threshold levels among the 3 indicators, we compared the error rates of pCI, pMDE, and pMDD with the different threshold levels in a pareto-plot displaying the false trust rate on one axis, and the false mistrust rate on the other. This method allows one to compare the different options simultaneously in both error rates and thus examine whether an indicator is pareto-superior to another, which means that we can find a threshold such that the indicator performs better than its alternatives in one of the error rates, while performing at least not worse in the other.

We provide an R Shiny App in the Supplemental Data and online (see *Data Accessibility Statement*), which allows one to specify experimental (sample size) and population (real effect size, variance, abundance) parameters to simulate and analyze similar experiments.

Results

Comparing the error rates of the 3 statistical indicators, we found that pCIs are pareto-superior over both pMDE and pMDD, irrespective of the threshold level (Figure 5B). For example, at a threshold level of 70%, pCI and pMDD produced comparable false mistrust rates (~20% of experiments with real zero effects had pCI and pMDD values larger than 70%, interpreted as "mistrust nonsignificant result"), but the false trust rate was much smaller for pCIs than for pMDDs (little more than

 © 2020 The Authors

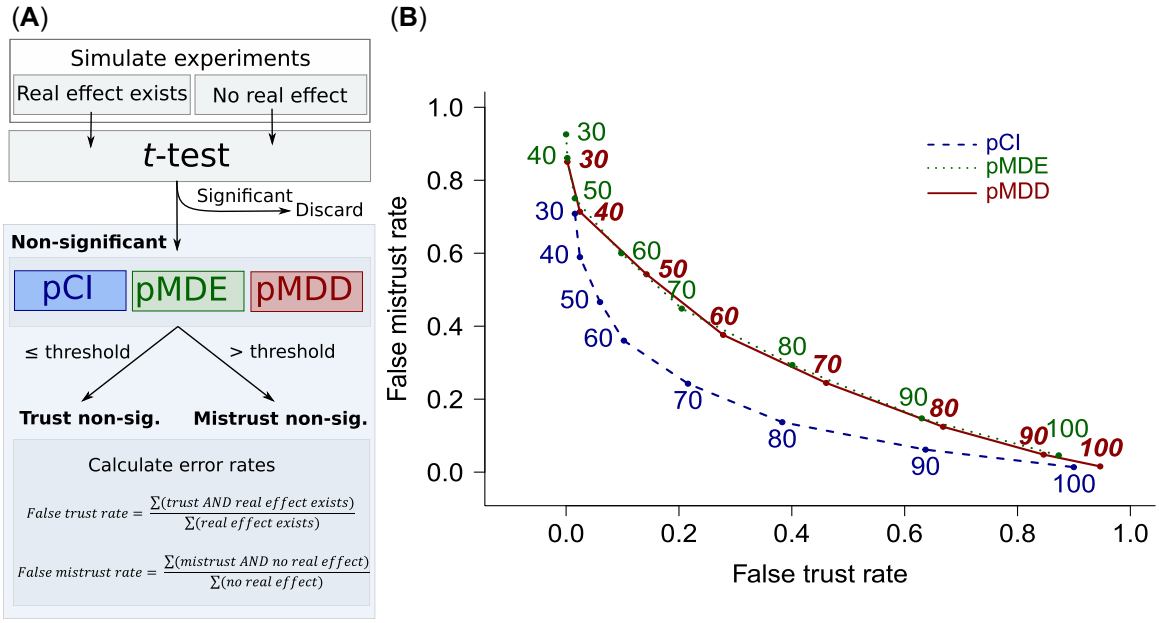

FIGURE 5: The translation of the continuous indicators proportional upper bound of the confidence interval (pCI), proportional minimum detectable effect (pMDE), and proportional minimum detectable difference (pMDD) into a dichotomous decision of trust/mistrust requires choosing a threshold level for the indicator. We can then examine the error rates of this decision (false trust/false mistrust rates) for the 3 methods and different thresholds. (**A**) Depiction of the entire procedure from the simulation of the data to the decisions about trust in a nonsignificant test outcome. (**B**) Rates at which pCI (blue, dashed), pMDE (green, dotted), and pMDD (red, solid; bold italics) erroneously suggest to trust (false trust rate) or mistrust (false mistrust rate) a nonsignificant result for varying threshold levels (numbers along lines). Rates are calculated from 1000 simulated experiments without effect (0 difference between means) and 1000 experiments with moderate effect (50% loss in mean abundance). Sample sizes (3–10), control means (10–100), and variance (i.e., theta values: 1–50) vary randomly among experiments. Values nearer to the bottom left are pareto-superior (i.e., more optimal) to values nearer to the top right. Numbers along lines indicate applied threshold levels at these points.

20% [pCI] and more than 40% [pMDD] of experiments with a real effect of 0.5, and nonsignificant test result were interpreted as "trust nonsignificant result").

Among themselves, pMDD and pMDE showed a virtually identical pattern. The only difference is that one has to choose different threshold levels for pMDD and pMDE to achieve the same false trust/false mistrust rates, which can be understood from the fact that MDD and MDE have different power (see section *Clarifying a Common Misunderstanding: The MDD Does Not Have Controlled Power*). Additional simulations show that these patterns hold for the detection of smaller (0.2) and larger (0.8) effects (see results in Figure S2, Supplemental Data 3.3). It is also independent of the chosen CI for the pCI. This only again amounts to a rescaling of the threshold: when the CI is changed, another threshold has to be set to achieve the same false trust/false mistrust rates (see results in Figure S3, Supplemental Data 3.4).

SENSITIVITY OF THE MDD AND OTHER SECONDARY FILTERS TO REAL EFFECT SIZE

One might argue that our previous test scenario, in which effect sizes are either 0 or large, is somewhat artificial. In reality, we might rather be faced with a situation in which effect sizes vary continuously. In this situation, rather than dividing 0 from non-zero effects, the regulatory goal might be more to distinguish small from large effects. This new scenario is both realistic and somewhat at odds with the base assumptions of the null hypothesis significance testing framework, which explicitly sets out to distinguish between zero and non-zero effects. Nevertheless, to examine whether the 3 indicators contribute additional information when applied post hoc on nonsignificant results in a situation in which true effects are practically never zero, but rather differ in size from small to large, we conducted a second set of simulated experiments. We investigated the correlation between real effect size of nonsignificant results and pCI, pMDD, and pMDE, and examined whether the 3 indicators differed in their ability to distinguish larger from smaller true effects.

Simulation of experiments

Data were generated identical to the previous simulations, except that real effect sizes were drawn randomly from a uniform distribution ranging from zero (no effect) to 1 (100% reduction in species abundance). Note that the continuous draws make an effect size of exactly zero infinitesimally unlikely, which means that we can assume that all effects are different from zero. Each simulated experiment was analyzed as before, followed by the calculation of pCI, pMDE, and pMDD. We then correlated pCI, pMDE, and pMDD with the real effect sizes both for all simulated experiments and only for experiments that yielded nonsignificant results ($p > 0.05$). In addition, nonsignificant experiments were filtered by applying different thresholds

© 2020 The Authors

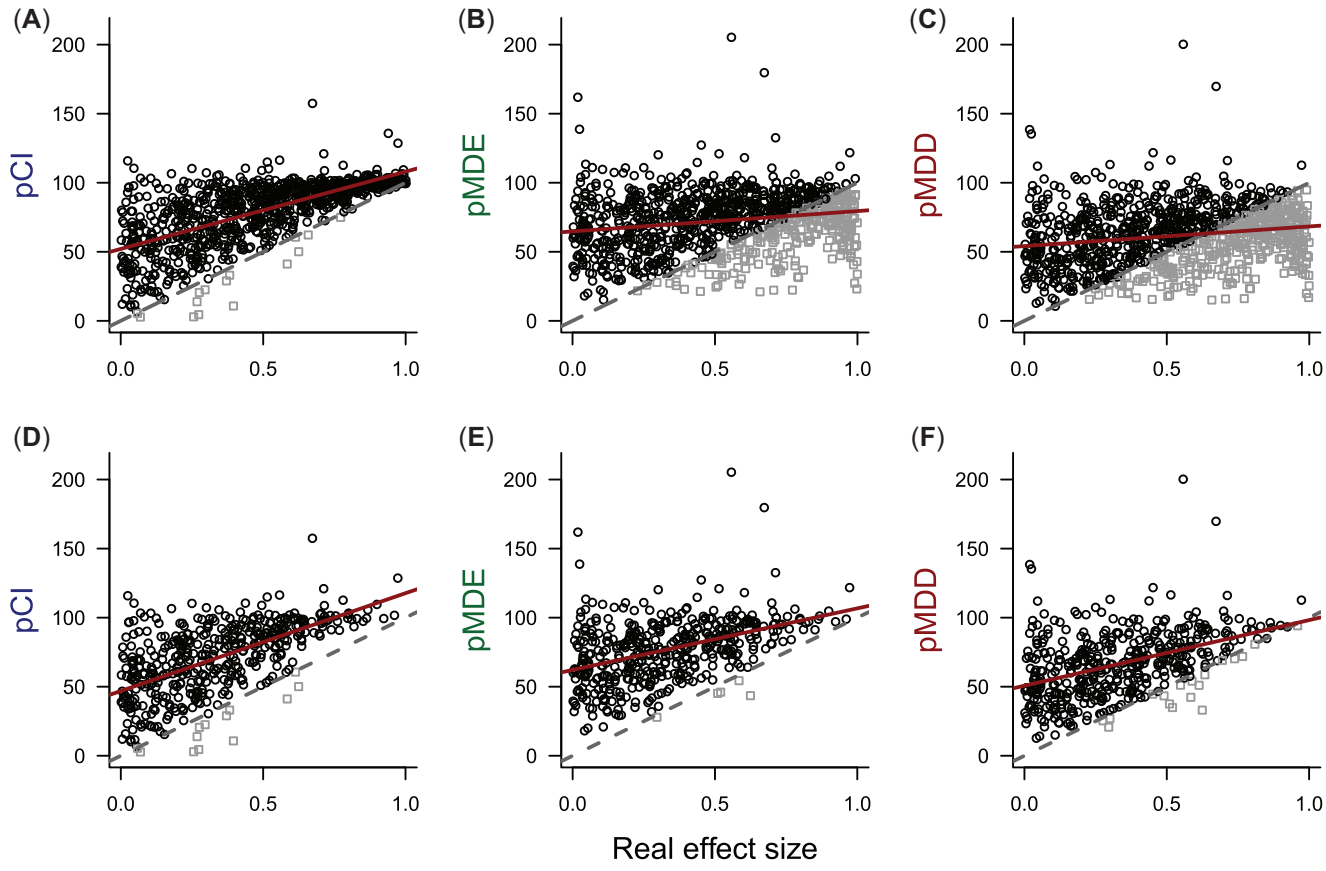

FIGURE 6: Correlation between proportional upper bound of the 95% confidence interval (pCI), (**A** and **D**), proportional minimum detectable effect with 80% power (pMDE), (**B** and **E**), and proportional minimum detectable difference (pMDD), (**C** and **F**) and real effect size for simulated experiments with known treatment effects ranging between 0 and 1 (0–100% reduction in species abundance). We show the values for (**A**–**C**) all simulated experiments including significant test results ($n = 1000$ experiments) and (**D**–**F**) only experiments that yielded nonsignificant results ($p > 0.05$; $n = 448$ experiments). Each data point represents one simulated experiment (i.e., one test). Solid line (red): regression line; dashed line (gray): pCI/pMDE/pMDD equals real effect; black dots: indicator was larger than the real effect; gray squares: indicator was lower than the real effect.

(range = 30–100, incremented by 1) on pCI, pMDE, and pMDD followed by comparison of the real effect sizes passing these secondary filters (R code in Supplemental Data 4). Similar simulations can be conducted via the R Shiny App provided in the *Supplemental Data* and online (see *Data Accessibility Statement*).

Results

Our simulations show that, in contrast to pCIs, neither pMDEs nor pMDDs correlate with real effect size when they are calculated on all results (Figure 6A–C). This is directly understandable from their mathematical definitions: the CI includes the effect size, whereas the MDE and MDD do not. In practice, however, single MDDs are in particular interpreted for nonsignificant results ($p > 0.05$). When only nonsignificant outcomes were considered, all indicators correlated with real effect size (Figure 6D–F) and all 3 indicators preferentially sorted out nonsignificant results with larger real effects (Figure 7). For all 3 indicators, real effect sizes passing these secondary filters were smaller compared with real effect sizes passing the *p* value filter alone (compare lines and boxplot in Figure 7: quantile lines after secondary filters are generally lower than quantiles of effect sizes after *p* value filter). With decreasing threshold levels, the numbers of larger real effect sizes passing the secondary filter also decrease (see decrease in quantile lines from left to right in Figure 7).

DISCUSSION

The MDD is a statistical indicator used in ecotoxicological risk assessment. Its main practical use case is similar to that of the MDE, which is commonly used in medicine and other scientific fields: to determine whether a nonsignificant result indicates the true absence of an effect rather than a lack of power. Our main results are that the MDD performs nearly identically to the MDE and is generally inferior to using CIs as a decision criterion for trust in a nonsignificant result. The reason is that MDD and MDE both calculate a "detectable" effect based on the variance of the data, independent of the actual estimated effect, whereas the CI includes the estimated effect size. Finally, we showed that the power of the MDD is variable,

 © 2020 The Authors

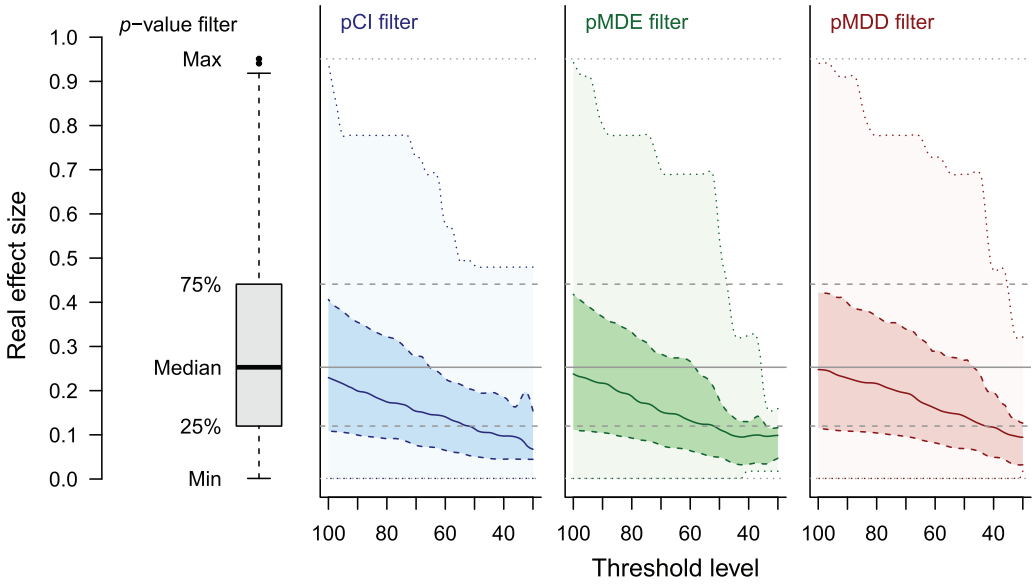

FIGURE 7: Effect sizes in experiments passing both the significance filter $p > 0.05$ (boxplot and gray lines in all plots, $n = 448$ experiments with nonsignificant result) and the secondary proportional upper bound of the 95% confidence interval (pCI), proportional minimum detectable effect with 80% power (pMDE), or proportional minimum detectable difference (pMDD) filter for different thresholds (horizontal axis) and different real effect sizes (vertical axis). Lines represent median (solid line), 25% and 75% quartiles (dashed lines), and minimum and maximum real effect size (dotted lines).

but generally substantially lower than 80%, and thus smaller than often assumed.

In more detail, our results show that MDD and MDE are virtually identical, and essentially only differ in their power, which is between 50 and 60% for the MDD, depending on the sample size and the distribution of the data and user-defined, but typically 80% for the MDE. These results are plausible from statistical theory when one considers that MDD and MDE differ only in that the former asks which effect size would have been significant, and the latter asks which effect size would have had a specified power. These differences can lead to slightly different numerical values when they are applied to finite sample sizes, but effectively, MDD and MDE are identical in all their properties, including their decisions about trust in nonsignificant results, as long as thresholds on pMDD and pMDE are rescaled to compensate for their different power.

Given this close relationship between MDE and MDD, it is not surprising that we identified many of the problems for the MDD that have been highlighted previously for MDEs as indicators for the true absence of effects (e.g., Hoenig and Heisey 2001; Colegrave and Ruxton 2003). The most important argument is that, instead of giving evidence for the presence or absence of real effects based on the estimated effect size, the MDE (and thus also the MDD) merely combine information about sample size and variance from the conducted experiment and translate this into a detectable effect. In other words, MDEs (and also MDDs) are identical if sample size and variance in the data are identical, regardless of the estimated effect size (Hoenig and Heisey 2001; Colegrave and Ruxton 2003).

Also in line with the statistical literature on the MDE (e.g., Hoenig and Heisey 2001), our simulations show that using CIs as a decision criterion for trust in nonsignificant results leads to lower error rates than with either MDEs and MDDs (Figure 4). The reason is that CIs combine both the variance and the estimated effect to decide whether to trust a nonsignificant result. In addition, when one is considering a continuous range of real effect sizes, as likely found in nature (Martínez-Abraín 2007), MDDs show no advantages over CIs as secondary effect size filters. The CIs have the additional advantage that they are readily available in most statistical software, and also when going beyond simple test procedures to more complicated regression models, for example, generalized linear mixed models (Bolker et al. 2009). We conclude that CIs are clearly preferable over MDDs and MDEs for the further interpretation of nonsignificant results, or for the interpretation of estimated effect sizes in general.

There are some alternatives to CIs that could also be considered in future research. Most importantly, there is the possibility of using equivalence tests instead of point null hypotheses. Equivalence tests test whether effects are significantly smaller than an effect size considered negligible or acceptable (McBride 1999; McBride et al. 2014; Harms and Lakens 2018; Lakens et al. 2018). By rejecting the null hypothesis in equivalence tests, we are able to state that effects are significantly lower than the chosen effect size, which circumvents the 2-step procedure. A systematic comparison of the error rates of equivalence tests with the error rates of confidence intervals would be interesting, but is beyond the scope of the present analysis, which essentially concentrated on the MDD, and procedures that work in the same analysis chain as the MDD. Another possibility would be the use of Bayes factors, which would allow one to directly compare the probability of H_1 (=effect size 0) with H_2

© 2020 The Authors

(=effect size different from 0), and thus provide the direct probability of an effect being present (Kass and Raftery 1995; Newman and Krull 2015).

Recommendations for practice

Based on our results, we make the following recommendations for ecotoxicological risk assessment on 2 levels.

If the statistical analysis of ecotoxicological risk assessments continues to rest on standard hypothesis tests with a point null hypothesis, CIs are clearly preferable over the MDD to control power and type II error rates. The MDD has a precise mathematical definition, and it is a perfectly valid indicator for what it measures, but it has clearly lower performance than the CI in distinguishing between the presence and absence of real effects. The CI is readily available for all statistical models, and for each pMDD threshold level, we can find a pCI threshold level that produces pareto-superior false trust/mistrust rates. If one is aiming at minimizing the error rates in recognizing false negatives among nonsignificant results, we therefore recommend the use of pCI thresholds instead of thresholds on the pMDD (a suggested procedure is summarized in Textbox 1).

Apart from the fact that a better alternative exists, we found that the MDD is often misinterpreted. For example, the MDD is often explained as the effect size that "can be detected" (European Food Safety Authority PPR Panel 2013; Brock et al. 2015; Andrade et al. 2017; Candolfi et al. 2018; European Food Safety Authority 2019) or "could have been detected" or "identified" (e.g., Peters et al. 2016; Rolke et al. 2016; Green et al. 2018) in the given experiment. These interpretations, however, neglect to indicate whether the effect size mentioned refers to the real effect in the population or the estimated effect from the experiment. A typical scientific reader would likely interpret these statements as suggesting that a real effect of the magnitude of the MDD could have been detected with high power. The underlying problem is that for the interpretation of the MDD, the differentiation between the real world (i.e., real effect size and real variance in the population) and the measured parameters (i.e., estimated effect size and estimated variance resulting from random sampling) is essential. The difficulty is that we want to get knowledge about the real world (real effect), but the MDD only provides information about an effect on the measurement level: for a given sample size and estimated variance, an estimated (i.e., measured) effect of the size of the MDD would be significant, always. A real effect size equal to the MDD, however, will be significant in only 50 to 60% of the conducted experiments (see *The MDD Does Not Have Controlled Power* section; Duquesne et al. 2020), which is far below the usually targeted power of 80% (European Food Safety Authority 2013; European Food Safety Authority PPR Panel 2015). The same differentiation has to be applied to the interpretation of the MDD as an upper bound for the real effect: for a nonsignificant experiment, the estimated effect will always be smaller than the MDD (information that is already covered by the nonsignificant *p* value; see also Hoenig and Heisey 2001; Colegrave and Ruxton 2003). What the MDD cannot tell us, however, is whether the real effect in the sampled population is also smaller than the MDD. Furthermore, in contrast to the CI, the rate at which the real effect is smaller the MDD is not fixed. Consequently, the MDD is not a reliable upper bound for a true effect (Hoenig and Heisey 2001; Colegrave and Ruxton 2003).

We consider that the MDD is probably no more difficult to explain and interpret than *p* values or CIs, and the latter 2 are often misinterpreted as well (Erickson and Rattner 2020), but we believe that the larger familiarity of the average reader with the CI constitutes another argument in its favor (see also Textbox 1, *Advantages of pCI over pMDD and pMDE*).

On a broader level, many of the issues discussed in this article fundamentally originate from the assumption that an ecotoxicological analysis must test for the absence of an effect in a null hypothesis significance testing framework, rather than estimating the posterior probability of an effect, or the size of an effect. Irrespective of the indicator (e.g., *p* value, pMDD, or pCI), a dichotomous decision will always require the choice of a threshold level. Translating such a threshold in well-interpretable probabilities has proved difficult for many statistical indicators, which is evidenced by the long-running discussions about the interpretation of *p* values (e.g., Cohen 1994; Greenland et al. 2016). Many researchers have therefore recommended a direct focus on estimating effect sizes and their precision (e.g., Newman 2008). This, however, would require regulators to define an acceptable level for the effect size, rather than requiring a "proof" of the absence of effects. Clearly, setting the levels for such acceptable impacts would be a difficult and politically sensitive challenge (Munkittrick et al. 2009; Mebane 2015). However, we note that applying thresholds on indicators, either on the pMDD or on the pCI, would effectively have the same effect. A direct discussion about acceptable effects, instead of an indirect setting via thresholds on the confidence for no effect, would avoid the definition of indicators that are difficult to interpret (e.g., MDDs) and would thereby likely increase the transparency of the regulatory procedure (Sanderson and Petersen 2002).

CONCLUSIONS

The MDD has a clear statistical definition, but its properties are often misunderstood. Most importantly, the MDD does not overcome the known problems of the minimum detectable effect (MDE) from post hoc power analysis. It is calculated based on the same principle, that is shifting the effect size to a specific point and differs from a standard post hoc power analysis (MDE) only in that it fixes the effect size to a particular *p* value, instead of a particular power. As a consequence of this, the power of the MDD is not controlled. More importantly, however, both MDD and MDE are defined independently of the estimated effect size, which makes them less powerful than CIs for detecting true negatives among nonsignificant results. Based on our findings, we recommend the use of CIs for the further interpretation of nonsignificant results. Moreover,

 © 2020 The Authors

TEXTBOX 1 Using confidence intervals for the interpretation of nonsignificant results

Recommended procedure

1. Perform null hypothesis significance test (in this example: one-sided *t* test for control mean larger than treatment mean) to get a *p* value for the comparison of the 2 groups. In mesocosm experiments, this would be the difference between the control group and one concentration level at one time point for one species.
2. Interpret significant outcomes ($p \leq 0.05$) as "real effect exists" (for this species at this concentration level and time point).
3. For nonsignificant outcomes: Calculate upper bound of the confidence interval (CI). For the *t* test, this can be calculated by:

$$CI = (mean_{control} - mean_{treatment}) + t_{critical} \times s_{res}\sqrt{\frac{1}{n_{control}} + \frac{1}{n_{treatment}}}$$

with s_{res} being the square root of the residual variance from a one-way analysis of variance (see Brock et al. 2015), *n* being the sample sizes of control and treatment, respectively, and $t_{critical}$ being the critical value of *t* (depending on α level and degrees of freedom).

4. Calculate pCI by relating CI upper bound to control mean (both back-transformed, if the test was conducted on transformed data):

$$pCI = \frac{CI_{upperbound}}{mean_{control}} \times 100.$$

5. Interpret the pCI: "The real effect is likely smaller than pCI." More precisely, the real effect is smaller than the upper bound of the pCI at a fixed rate that is in 95% (i.e., the *confidence level*) of performed tests (CI calculated for one-sided test).
6. Apply threshold level on pCI, if a dichotomous decision is necessary, and interpret the result within the context of additional tests for different time points and concentration levels (temporal pattern, dose-response relationship).

Example calculation (data taken from Brock et al. 2015)

Original data:

Replicate	Control	Treatment
1	175	29
2	65	114
3	154	72
4	83	—
Mean	119.25	71.67

*ln(2*x + 1)*-transformed data:

Replicate	Control	Treatment
1	5.86	4.08
2	4.88	5.43
3	5.73	4.98
4	5.12	—
Mean	5.40	4.83

p-value (one-sided *t* test, control mean larger treatment mean, equal variances):	0.124
CI upper bound (95% confidence level, one-sided, transformed level):	1.366
Back-transformed CI upper bound (for calculation see Brock et al. 2015):	82.25
pCI (=back-transformed CI upper bound/back-transformed control mean × 100):	74.8%

Interpretation of the CI:

The real effect is smaller than the pCI (with an error rate of 5%). Thus, in the present study, the real reduction in species abundance in the treatment as compared to the control (for this species at this time point) was likely smaller than 74.8%.

(*Continued*)

© 2020 The Authors

(*Continued*)

For comparison:

MDD (transformed level):	0.80
Back-transformed MDD (for calculation see Brock et al. 2015):	60.88
pMDD (=back-transformed MDD/back-transformed control mean × 100):	55.42%

Interpretation of the MDD:
In this example experiment, a reduction in species abundance of 55.42% as compared to the control would have been significant, if measured.

Advantages of pCI over pMDD and pMDE

1. pCIs produce more beneficial error rates in decisions about the presence or absence of real effects compared to pMDDs and pMDEs (see Figure 5).
2. Among 2 nonsignificant tests with equal variances, pCIs are able to identify the test with lower estimated effect size as stronger evidence for the true absence of an effect, which is not the case for pMDDs and pMDEs (see Figure 2).
3. pCIs can be readily obtained from statistical software, they may be more familiar to many analysts (and thus better understood), and their interpretation as an upper bound for the real effect with fixed error rate (present study: real effect is lower than pCI in 95% of experiments) seems well-communicable also to general decision-makers.

considering the low likelihood of complete zero effects, we argue that a stronger emphasis on estimated effect size and their CIs instead of hypothesis tests would lead to better decisions and greater transparency in risk assessments.

Supplemental Data—The Supplemental Data are available on the Wiley Online Library at https://doi.org/10.1002/etc.4847.

Acknowledgment—We are grateful to J.-D. Ludwigs, T. Pamminger, A. Singer, and 2 anonymous reviewers for commenting on an earlier version of the manuscript. M.M. Mair is supported by the Christiane Nüsslein-Volhard Foundation. Open access funding enabled and organized by Projekt DEAL.

This article has earned an Open Data/Materials badge for making publicly available the digitally shareable data necessary to reproduce the reported results. Both the data and materials are available at https://github.com/TheoreticalEcology/Mair-et-al-2020 and https://doi.org/10.5281/zenodo.3951619. In addition, an R Shiny-app for running the code of the experimental simulations can be found at https://mair-et-al-2020.shinyapps.io/mddshiny/. Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki.

Data Availability Statement—All R code pertaining to this manuscript is provided in the Supplemental Data. In addition, an R Shiny-app for running the code of the experimental simulations can be found on https://mair-et-al-2020.shinyapps.io/mddshiny/. All materials can also be found at https://doi.org/10.5281/zenodo.3951619 (https://github.com/TheoreticalEcology/Mair-et-al-2020).

REFERENCES

Andrade TO, Bergtold M, Kabouw P. 2017. Minimum significant differences (MSD) in earthworm field studies evaluating potential effects of plant protection products. *J Soils Sediments* 17:1706–1714.

Baguley T. 2004. Understanding statistical power in the context of applied research. *Appl Ergon* 35:73–80.

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White J-SS. 2009. Generalized linear mixed models: A practical guide for ecology and evolution. *Trends Ecol Evol* 24:127–135.

Brock TCM, Hammers-Wirtz M, Hommen U, Preuss TG, Ratte H-T, Roessink I, Strauss T, van den Brink PJ. 2015. The minimum detectable difference (MDD) and the interpretation of treatment-related effects of pesticides in experimental ecosystems. *Environ Sci Pollut Res* 22:1160–1174.

Brock TCM, Lahr J, van den Brink PJ. 2000. Ecological risks of pesticides in freshwater ecosystems; Part 1: Herbicides. Alterra-Rapport No 88. Alterra, Green World Research, Wageningen, The Netherlands.

Cabrera AR, Almanza MT, Cutler GC, Fischer DL, Hinarejos S, Lewis G, Nigro D, Olmstead A, Overmyer J, Potter DA, Raine NE, Stanley-Stahr C, Thompson H, van der Steen J. 2016. Initial recommendations for higher-tier risk assessment protocols for bumble bees, *Bombus* spp. (Hymenoptera: Apidae). *Integr Environ Assess Manag* 12:222–229.

Candolfi MP, Bargen H, Bocksch S, Klein O, Kleinhenz M, Knaebe S, Szczesniak B. 2018. Which endpoints can be reliably assessed in semi-field pollinator species testing without estimating false positive or false negative? MDD's and replicates issue. *J Agric Sci Technol A* 8:142–161.

Cohen J. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum, Mahwah, NJ, USA.

Cohen J. 1994. The earth is round ($p < .05$). *Am Psychol* 49:997–1003.

Colegrave N, Ruxton GD. 2003. Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. *Behav Ecol* 14:446–447.

de Jong F, Brock TCM, Foekema EM, Leeuwangh P. 2008. Guidance for summarizing and evaluating aquatic micro- and mesocosm studies. RIVM Report 601506009/2008. RIVM, Bilthoven, The Netherlands.

de Jong F, van Beelen P, Smit C, Montforts M. 2006. Guidance for summarising earthworm field studies. RIVM Report 601506006/2006. RIVM, Bilthoven, The Netherlands.

Duquesne S, Alalouni U, Gräff T, Frische T, Pieper S, Egerer S, Gergs R, Wogram J. 2020. Better define beta–optimizing MDD (minimum

 © 2020 The Authors

detectable difference) when interpreting treatment-related effects of pesticides in semi-field and field studies. *Environ Sci Pollut Res*. https://doi.org/10.1007/s11356-020-07761-0

Environment Canada. 2005. Guidance document on statistical methods. Ottawa, ON, Canada.

Erickson RA, Rattner BA. 2020. Moving beyond $p < 0.05$ in ecotoxicology: A guide for practitioners. *Environ Toxicol Chem*. https://doi.org/10.1002/etc.4800

European Commission. 2009. Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market and repealing Council Directives 79/117/EEC and 91/414/EEC. *OJ* L309/1, 24.11:1–50.

European Commission. 2020. Eurostat. Pesticide sales [aei_fm_salpest09]. Luxembourg City, Luxembourg. [cited 2020 May 21]. Available from: http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=aei_fm_salpest09&lang=en

European Food Safety Authority. 2013. Guidance document on the risk assessment of plant protection products on bees (*Apis mellifera*, *Bombus* spp. and solitary bees). *EFSA J* 11:3295.

European Food Safety Authority. 2019. Outcome of the pesticides peer review meeting on general recurring issues in ecotoxicology. EN-1673. Parma, Italy.

European Food Safety Authority PPR Panel. 2013. Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters. *EFSA J* 11:3290.

European Food Safety Authority PPR Panel. 2015. Technical report on the outcome of the pesticides peer review meeting on general recurring issues in ecotoxicology. EN-924. Parma, Italy.

European Food Safety Authority PPR Panel. 2017. Scientific opinion addressing the state of the science on risk assessment of plant protection products for in-soil organisms. *EFSA J* 15:4690.

Fagley NS. 1985. Applied statistical power analysis and the interpretation of nonsignificant results by research consumers. *J Couns Psychol* 32:391–396.

Food and Agriculture Organization of the United Nations. 2020. FAOstat. Pesticides use. Rome, Italy. [cited 2020 May 21]. Available from: http://www.fao.org/faostat/en/#data/RP

Fox DR, Landis WG. 2016. Don't be fooled-A no-observed-effect concentration is no substitute for a poor concentration-response experiment. *Environ Toxicol Chem* 35:2141–2148.

Goodman SN, Berlin JA. 1994. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med* 121:200–206.

Green JW, Springer TA, Holbech HH. 2018. *Statistical Analysis of Ecotoxicity Studies*. John Wiley & Sons, Hoboken, NJ, USA, pp 310–313.

Greenland S. 2012. Nonsignificance plus high power does not imply support for the null over the alternative. *Ann Epidemiol* 22:364–368.

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. 2016. Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *Eur J Epidemiol* 31:337–350.

Harms C, Lakens D. 2018. Making "null effects" informative: Statistical techniques and inferential frameworks. *J Clin Transl Res* 3(Suppl 2): 382–393.

Hoenig JM, Heisey DM. 2001. The abuse of power. *Am Stat* 55:19–24.

Höss S, Reiff N, Nguyen HT, Jehle JA, Hermes H, Traunspurger W. 2014. Small-scale microcosms to detect chemical induced changes in soil nematode communities—Effects of crystal proteins and Bt-maize plant material. *Sci Total Environ* 472:662–671.

Jensen JLWV. 1906. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Math* 30:175–193.

Johnson NL, Welch BL. 1940. Applications of the non-central t-distribution. *Biometrika* 31:362–389.

Johnson PCD, Barry SJE, Ferguson HM, Müller P. 2015. Power analysis for generalized linear mixed models in ecology and evolution. *Methods Ecol Evol* 6:133–142.

Kass RE, Raftery AE. 1995. Bayes factors. *J Am Stat Assoc* 430:773–795.

Kennedy JH, Ammann LP, Waller WT, Warren JE, Hosmer AJ, Cairns SH, Johnson PC, Graney RL. 1999. Using statistical power to optimize sensitivity of analysis of variance designs for microcosms and mesocosms. *Environ Toxicol Chem* 18:113–117.

Kraufvelin P. 1998. Model ecosystem replicability challenged by the "soft" reality of a hard bottom mesocosm. *J Exp Mar Biol Ecol* 222:247–267.

Lakens D, Scheel AM, Isager PM. 2018. Equivalence testing for psychological research: A tutorial. *Adv Methods Pract Psychol Sci* 1:256–269.

Laskowski R. 1995. Some good reasons to ban the use of NOEC, LOEC and related concepts in ecotoxicology. *Oikos* 73:140–144.

Lehmann R, Bachmann J, Maletzki D, Polleichtner C, Ratte HT, Ratte M. 2016. A new approach to overcome shortcomings with multiple testing of reproduction data in ecotoxicology. *Stoch Environ Res Risk Assess* 30:871–882.

Lemoine NP, Hoffman A, Felton AJ, Baur L, Chaves F, Gray J, Yu Q, Smith MD. 2016. Underappreciated problems of low replication in ecological field studies. *Ecology* 97:2554–2561.

Lenth RV. 2001. Some practical guidelines for effective sample size determination. *Am Stat* 55:187–193.

Liber K, Kaushik NK, Solomon KR, Carey JH. 1992. Experimental designs for aquatic mesocosm studies: A comparison of the "anova" and "regression" design for assessing the impact of tetrachlorophenol on zooplankton populations in limnocorrals. *Environ Toxicol Chem* 11:61–77.

Martínez-Abraín A. 2007. Are there any differences? A non-sensical question in ecology. *Acta Oecologica* 32:203–206.

McBride G, Cole RG, Westbrooke I, Jowett I. 2014. Assessing environmentally significant effects: A better strength-of-evidence than a single *P* value? *Environ Monit Assess* 186:2729–2740.

McBride GB. 1999. Equivalence tests can enhance environmental science and management. *Aust NZ J Stat* 41:19–29.

Mebane C. 2015. In response: Biological arguments for selecting effect sizes in ecological testing—A governmental perspective. *Environ Toxicol Chem* 34:2440–2442.

Munkittrick KR, Arens CJ, Lowell RB, Kaminski GP. 2009. A review of potential methods of determining critical effect size for designing environmental monitoring programs. *Environ Toxicol Chem* 28:1361–1371.

Newman MC. 2008. "What exactly are you inferring?" A closer look at hypothesis testing. *Environ Toxicol Chem* 27:1013–1019.

Newman MC, Krull M. 2015. In response: Regression or significance tests: What other choice is there?—An academic perspective. *Environ Toxicol Chem* 34:2439–2440.

Nys C, Van Regenmortel T, De Schamphelaere K. 2019. The effects of nickel on the structure and functioning of a freshwater plankton community under high dissolved organic carbon conditions: A microcosm experiment. *Environ Toxicol Chem* 38:1923–1939.

Organisation for Economic Co-operation and Development. 1998. Report of the OECD workshop on statistical analysis of aquatic toxicity data. *Series on Testing and Assessment*, No 10. ENV/MC/CHEM(98)18. Paris, France.

Organisation for Economic Co-operation and Development. 2006. Guidance document on simulated freshwater lentic field tests (outdoor microcosms and mesocosms). *Series on Testing and Assessment*, No 53. ENV/JM/MONO(2006)17. Paris, France.

Organisation for Economic Co-operation and Development. 2007. Guidance document on the honeybee (*Apis mellifera* L.) brood test under semi-field conditions. *Series on Testing and Assessment*, No 75. ENV/JM/MONO(2007)22. Paris, France.

O'Keefe DJ. 2007. Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of statistical power analyses. *Commun Methods Meas* 1:291–299.

Onwuegbuzie AJ, Leech NL. 2004. Post hoc power: A concept whose time has come. *Underst Stat* 3:201–230.

Owen DB. 1965. The power of Student's t-test. *J Am Stat Assoc* 60:320–333.

Peters B, Gao Z, Zumkier U. 2016. Large-scale monitoring of effects of clothianidin-dressed oilseed rape seeds on pollinating insects in Northern Germany: Effects on red mason bees (*Osmia bicornis*). *Ecotoxicology* 25:1679–1690.

Rolke D, Persigehl M, Peters B, Sterk G, Blenau W. 2016. Large-scale monitoring of effects of clothianidin-dressed oilseed rape seeds on pollinating insects in Northern Germany: Residues of clothianidin in pollen, nectar and honey. *Ecotoxicology* 25:1691–1701.

SANCO. 2002a. Guidance document on terrestrial ecotoxicology under council directive 91/414Draft Working Document. SANCO/1039/2002 rev 2 final. European Union (DG Health and Consumer Protection), Brussels, Belgium.

SANCO. 2002b. Guidance document on aquatic ecotoxicology in the context of the Directive 91/414/EEC. Working Document. SANCO/3268/2001 rev.4 (final). European Union (DG Health and Consumer Protection), Brussels, Belgium.

© 2020 The Authors

Sanderson H. 2002. Pesticide studies. *Environ Sci Pollut Res* 9:429–435.

Sanderson H, Petersen S. 2002. Power analysis as a reflexive scientific tool for interpretation and implementation of the precautionary principle in the European Union. *Environ Sci Pollut Res* 9:221–226.

Scholz-Starke B, Beylich A, Moser T, Nikolakis A, Rumpler N, Schäffer A, Theißen B, Toschki A, Roß-Nickoll M. 2013. The response of soil organism communities to the application of the insecticide lindane in terrestrial model ecosystems. *Ecotoxicology* 22:339–362.

Steidl RJ, Hayes JP, Schauber E. 1997. Statistical power analysis in wildlife research. *J Wildl Manag* 61:270–279.

Szöcs E, Schäfer RB. 2015. Ecotoxicology is not normal. *Environ Sci Pollut Res* 22:13990–13999.

Tilman D, Cassman KG, Matson PA, Naylor R, Polasky S. 2002. Agricultural sustainability and intensive production practices. *Nature* 418:671–677.

Touart LW. 1994. Regulatory endpoints and the experimental design of aquatic mesocosm tests. In Graney RL, Kennedy JH, Rodgers JH, eds, *Aquatic Mesocosm Studies in Ecological Risk Assessment*. CRC, Boca Raton, FL, USA, pp 25–33.

van Dam RA, Harford AJ, Warne MSJ. 2012. Time to get off the fence: The need for definitive international guidance on statistical analysis of ecotoxicity data. *Integr Environ Assess Manag* 8:242–245.

van der Hoeven N. 2008. Calculation of the minimum significant difference at the NOEC using a non-parametric test. *Ecotoxicol Environ Saf* 70:61–66.

Williams DA. 1972. The comparison of several dose levels with a zero dose control. *Biometrics* 28:519–531.

Zumbo BD, Hubley AM. 1998. A note on misconceptions concerning prospective and retrospective power. *Statistician* 47:385–388.

 © 2020 The Authors