Environ Sci Pollut Res
DOI 10.1007/s11356-013-1749-z

RESEARCH ARTICLE

Misuse of null hypothesis significance testing: would estimation of positive and negative predictive values improve certainty of chemical risk assessment?

Mirco Bundschuh · Michael C. Newman · Jochen P. Zubrod · Frank Seitz · Ricki R. Rosenfeldt · Ralf Schulz

Received: 21 February 2013 / Accepted: 15 April 2013

Abstract Although generally misunderstood, the *p* value is the probability of the test results or more extreme results given H_0 is true: it is not the probability of H_0 being true given the results. To obtain directly useful insight about H_0, the positive predictive value (PPV) and the negative predictive value (NPV) may be useful extensions of null hypothesis significance testing (NHST). They provide information about the probability of statistically significant and non-significant test outcomes being true based on an a priori defined biologically meaningful effect size. The present study explores the utility of PPV and NPV in an ecotoxicological context by using the frequently applied *Daphnia magna* reproduction test (OECD guideline 211) and the chemical stressor lindane as a model system. The results indicate that especially the NPV deviates meaningfully between a test design strictly following the guideline and an experimental procedure controlling for α and β at the level of 0.05. Consequently, PPV and NPV may be useful supplements to NHST that inform the researcher about the level of confidence warranted by both statistically significant and non-significant test results. This approach also reinforces the value of considering α, β, and a biologically meaningful effect size a priori.

Keywords Sample size · Bayesian · Power analysis · Effect size · Type I error rate · Type II error rate

Responsible editor: Philippe Garrigues

M. Bundschuh (✉) · J. P. Zubrod · F. Seitz · R. R. Rosenfeldt · R. Schulz
Institute for Environmental Sciences,
University of Koblenz-Landau, Fortstrasse 7,
76829 Landau, Germany
e-mail: bundschuh@uni-landau.de

M. Bundschuh
e-mail: mirco.bundschuh@slu.se

M. Bundschuh
Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden

M. C. Newman
Virginia Institute of Marine Science, College of William and Mary, Virginia, USA

Introduction

Null hypothesis significance testing (NHST) is an inferential tool used *inter alia* in the fields of biology, ecotoxicology, and environmental risk assessment of chemicals (Newman 2008, 2013). At the same time, the misinterpretation of NHST-derived *p* values is pervasive, occasionally leading to incorrect inferences. For instance, Gigerenzer (2004) performed a survey among students and lecturers from different psychology departments demonstrating that only 20 % of statistics teachers and none of their students were able to identify the correct interpretation of the NHST *p* value. This is the probability of the obtained or even more extreme results, given the null hypothesis of no effect is true (i.e., $p(D|H_0)$), not the more directly relevant probability of the null hypothesis being true given the results (i.e., $p(H_0|D)$) (for further reading see Newman 2008, 2013). Likewise, Newman (2013) performed a comparable survey among environmental scientists at five occasions and obtained an equally alarming outcome.

In this context, the application of the NHST-based no observed effect concentration (NOEC) or the lowest observed effect concentration (LOEC) in the fields of ecotoxicology and environmental risk assessment is, although still representing

Published online: 28 April 2013

the foundation of many standard test protocols (e.g., OECD 2008), frequently criticized (e.g., Jager 2012; Landis and Chapman 2011). The NOEC is defined as the highest treatment concentration in a toxicity experiment not being statistically significantly different from the uncontaminated control, while the LOEC is the first treatment concentration that is statistically significantly different from the control. This concept, however, has several shortcomings in addition to the above-mentioned general misinterpretation of NHST (e.g., Fox 2009). Firstly, only statistical significance is taken into consideration and any statistically non-significant difference from the control is immaterial. However, a lack of statistical significance is not equivalent to a lack of biological significance or environmental concern, which should be defined a priori on the basis of expert knowledge (Crane and Newman 2000). Secondly, this dichotomous decision—adverse effect vs. no adverse effect—is highly influenced by study design and the associated type I (α) and II (β) error rates (van der Hoeven 1998). Usually the type I error rate reflecting the probability of falsely rejecting the null hypothesis of no difference is set at $\alpha<0.05$. The type I error rate is also maintained at this level with experimental designs involving multiple statistical comparisons as is typical of conventional experiments that produce NOEC/LOEC estimates. The probability of falsely rejecting the null hypothesis (β) remains ill-defined (Mudge et al. 2012) although statistical tools to estimate power ($1-\beta$) could easily be applied to define β prior to experimentation.

As power analysis for any specific test uses the relationship of sample size, α, effect size (ES) and associated variability as well as the desired β (Nakagawa and Forster 2004), it can also be applied to estimate the minimal adequate sample size of a planned experiment predicated on the α, β, and ES being set based on defensible expert judgment and a priori knowledge. Thus, it allows control of β and consequently statistical power (Nakagawa and Cuthill 2007). Since power analysis is customarily given short shrift, difficulties may arise during testing given a low, but unknown statistical power, e.g., the statistical power of the OECD *Daphnia magna* reproduction test (OECD 2008), is as low as 0.8 ($\beta=0.2$) for an ES of approximately 30 % (van der Hoeven 1998). Further standardized reproduction experiments using *Eisenia fetida* or *Folsomia candida* exhibit a statistical power of approximately 0.5 at a comparable ES (van der Hoeven 1998). This is also the case for experiments in the fields of behavioral ecology and animal behavior (Nakagawa and Cuthill 2007), and therefore probably also higher tier studies in ecotoxicology, e.g., mesocosms. The latter are occasionally requested during the environmental risk assessment of chemicals. The power to detect a genuine difference with NHST in such cases might be similar to that of flipping a fair coin (Nakagawa and Cuthill 2007). Putting this in the context of chemical risk assessment, a fixed α of 0.05 combined with a dubiously high β favors an incorrect decision that an unsafe concentration of a given substance is safe. Consequently, balancing α and β, or taking the latter more seriously, is recommended (Newman 2008, 2013)—especially during environmental risk assessment of chemical stressors. Additional NHST misuses are discussed in the literature (e.g., Ioannidis 2005; Kline 2004; Nakagawa and Cuthill 2007; Newman 2008) but are beyond the scope of the present study.

Several alternatives, which are nonetheless based on testable hypothesis, are suggested to avoid the described shortcomings of NHST: The confidence interval approach is, for instance, frequently applied in the field of psychology (Kline 2004) and was also proposed for ecotoxicological investigations (Newman 2008, 2013). It provides a point estimate of an ES together with the precision of the measurement. If a 95 % confidence interval around this point estimate is chosen as a measure of precision, which is comparable to α set at 0.05 in NHST, inferences about statistical significance can be made (Altman et al. 2000; Bundschuh et al. 2011; Newman 2008). Additionally, multiple significance levels can be displayed simultaneously (Zubrod et al. 2011). Dose–response modeling is another suitable approach, which can bypass many of the issues related to NHST and predict concentrations causing x % effect in the response variable following a given time of exposure (Landis and Chapman 2011). On the basis of dose–response models, receiver operator characteristic (ROC) curves can be obtained. ROC curves inform about the probability of predicting an effect falsely positive and falsely negative at a given concentration (Newman 2013). Also, a Bayesian method has been proposed, which determines no-effect exposure concentrations based on dose–response data sets taking a priori knowledge into account (Fox 2010).

In the context of "Bayesian-like" statistics, the positive predictive value (PPV) and the negative predictive value (NPV) are suggested as one of several useful extensions of or alternatives to NHST (Ioannidis 2005; Newman 2008). However, the PPV and NPV should be seen as the first step—and not as the ultimate alternative method—going beyond the routine application of NHST. A PPV of 0.6, for instance, indicates a 60 % post study probability of the alternative hypothesis (H_A) being true given a statistically significant outcome (Ioannidis 2005). In other words, the probability of a biologically meaningful ES exceeding an a priori set ES is 60 % with a statistically significant test result. Likewise, the NPV provides the estimated probability for the ES being below the a priori defined ES given a statistically non-significant test outcome (Ioannidis 2005).

Although these methods are commonplace in other fields such as medical diagnostics (Ioannidis 2005), they are generally ignored in the field of ecotoxicology and the environmental risk assessment of chemicals, which may result in incorrect conclusions even though the demand for more definitive inferences is increasing. Therefore, this manuscript demonstrates using the frequently applied OECD *D. magna* reproduction test (OECD 2008) and the chemical stressor lindane as a model system that inferentially valid conclusions can be reached if attention is paid to predictive instead of *p* values.

Material and methods

Obtaining prior knowledge

Prior to the preliminary experiment, a literature search was performed in April 2012 using both ISI Web of Knowledge and the ECOTOX database provided by the US Environmental Protection Agency. This literature search produced three studies of lindane effects on *D. magna* reproduction, which were published between 1995 and 2004. The LOECs ranged from 100 to 250 μg/L and the ES at these concentrations were between 25 and 60 % (Antunes et al. 2004; DeCoen and Janssen 1997; Ferrando et al. 1995). This information was applied to select an appropriate range of test concentrations (preliminary experiment: 0, 50, 100, 200, 400, 800 μg/L; definitive experiment: 0, 100, 200, 300, 400, 500 μg/L) and to estimate the a priori probability (*R*), for each concentration investigated (see below). It should be mentioned that *R* is not necessarily based on studies that investigated exactly the same test organism or endpoint. Although increasing the uncertainty associated with the calculation of PPV and NPV considerably, it may also be adequate to consider other related species and response variables.

Validation of the test performance and determination of treatment variability

D. magna (clone V) were cultured in-house at 20±1 °C with a 16:8-h (light/dark) photoperiod in reconstituted hard freshwater (ASTM 2007) enriched with selenium, vitamins (thiamine hydrochloride, cyanocobalamine, biotine), and seaweed extract (Marinure®, Glenside, Scotland). Daphnids were fed the green algae *Desmodesmus* sp. on a daily basis (~200 μg C per organism). Lindane (Fluka, Germany) was applied as analytical standard dissolved in acetone (purity >99.9 %, Roth, Germany), which necessitated a solvent control in the experimental design containing 0.1 % solvent. Because no difference between solvent control and control was observed (results not shown), only the control was used for statistical assessments. In general, the reproduction experiments followed the recommendations of the OECD guideline 211 (OECD 2008) with the endpoint being the mean number of offspring per adult.

A preliminary reproduction experiment was performed following the guideline strictly, to validate the test performance by comparing the outcome with the literature. A Dunnett's test for multiple comparisons indicated a LOEC of 400 μg/L, which is approximately twice as high as the values reported in literature but still in general agreement (Fig. 1).

The preliminary experiment also provided estimates of the response variability in terms of the standard deviation (SD) needed for the power analysis. Although it would introduce uncertainty into the calculations, this variability might also be derived from experiments performed earlier with other environmental stressors, rendering preliminary experiments unnecessary. Albeit this approach requires homogeneity of variances among treatments, even if the requirement is met, the variability associated with the treatment means vary slightly from each other. Hence, it is recommended that variability of single treatments is taken into account, assuming that variability is not concentration dependent, but can differ randomly in a certain range. In the context of the present study, the standard deviation of the preliminary experiment's control treatment (SD_c) and the one from the treatment exhibiting the highest variability (SD_i), i.e., the 400-μg/L treatment, were pooled to produce

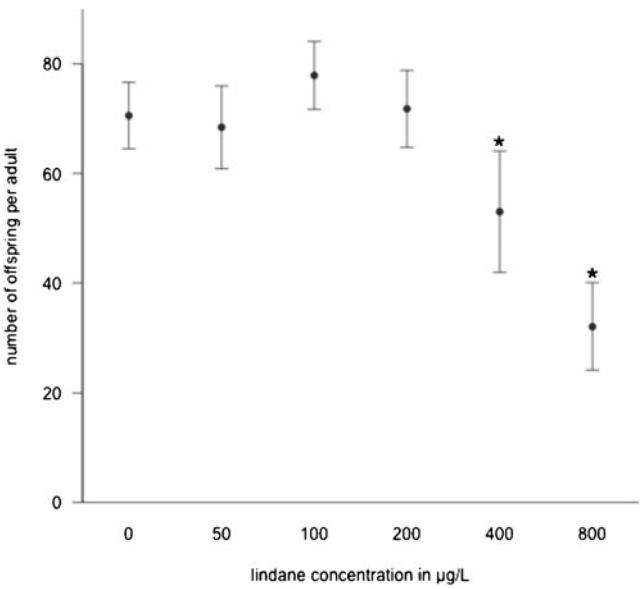

Fig. 1 Mean (±95 % confidence intervals; $n=10$) number of offspring released per adult *D. magna* at different concentrations of lindane during the preliminary experiment. *Asterisks* (*) denote statistically significant differences (experimentwise $\alpha=0.05$) compared to the control with a Dunnett's multiple comparisons test

one standard deviation (SD_{pooled}) for the power analysis (Altman et al. 2000) as follows:

$$SD_{pooled} = \sqrt{\frac{SD_c^2 + SD_i^2}{2}} \quad (1)$$

The described procedure considers the worst case scenario in terms of variability associated with the investigated endpoint. Hence, the statistical power obtained can be considered as a conservative estimate.

Power analysis

Power analysis for a two-sided Dunnett's test was accomplished to estimate the minimal adequate sample size per treatment (van der Hoeven 1998). As an impairment in *D. magna* reproduction of 25 % might translate into impaired population development (e.g., Preuss et al. 2010), this level of effect was chosen as a biologically significant ES during power analysis of the present study. Another threshold might have been selected depending on the endpoint and species investigated as well as on the basis of further ecological knowledge and scientific experience. The variability of the data was estimated with SD_{pooled}, and α and β were fixed at 0.05 to reflect our judgment that both error rates were equally serious. The power analysis for a two-sided Dunnett's test with five comparisons suggested a minimal sample size of 32 for the control and 14 for each of the remaining treatments. This difference in sample sizes for control (n_c) and the remaining treatments (n_i) is due to the optimal allocation of replicates in terms of statistical power, which depends on the number of treatments (q) investigated (Dunnett 1955):

$$\frac{n_c}{n_i} = \sqrt{q-1} \quad (2)$$

Calculation of PPV and NPV

In comparison to many Bayesian metrics, the PPV and NPV require instead of the specification of an a priori probability distribution, the less challenging a priori probability (R), which can be estimated from existing literature (Wacholder et al. 2004). It reflects the proportion of statistically significant observations in studies done prior to the current one at each treatment (=concentration of the tested compound; see also Obtaining prior knowledge and Validation of the test performance and determination of treatment variability):

$$R = \frac{\text{Number of studies reporting statistically significant effects}}{\text{Total number of studies considered}} \quad (3)$$

Moreover, R may also be estimated based on expert judgment considering environmental chemistry and toxicology data and would hence be independent from the availability of data for a particular test species (Newman 2008). However, considering criticisms (e.g., the strong dependences on statistical power) of the NOEC/LOEC concept (Fox 2009; Jager 2012; Landis and Chapman 2011), it is equivocal whether or not it would be worthwhile considering the ES reported in each of the studies (potentially also weighting the study according to the respective precision) instead of statistically significant deviations during calculation of R. The proportion of studies reporting an ES that exceeded a pre-defined threshold-ES from the total number of published studies on the endpoint investigated might be used instead. Indeed, this suggestion would limit the confounding effects introduced into the calculations of the PPV and NPV by the NOEC/LOEC concept, which is particularly evident in the study of Ferrando et al. (1995) suggesting an ES of 60 % as a low effect size. In the present study, the threshold ES was set at 25 % reduction in the number of offspring released per adult. Given this threshold, the manner by which R is calculated makes little difference

Table 1 α, ES, β, n_c, n_i, R, PPV as well as NPV are displayed for each concentration investigated in the definitive (=adapted) experiment and for a test design strictly following the OECD guideline 211

	Lindane concentration (µg/L)									
	100		200		300		400		500	
Test design	Adapted	OECD	Adapted	OECD	Adapted	OECD	Adapted	OECD	Adapted	OECD
α	0.05									
ES	25 %									
β	0.05	0.20	0.05	0.20	0.05	0.20	0.05	0.20	0.05	0.20
n_c/n_i	32/14	10/10	32/14	10/10	32/14	10/10	32/14	10/10	32/14	10/10
R	0.25		0.5		0.75		1.00		1.00	
PPV (%)	82.61	80.00	90.48	88.89	93.44	92.31	95.00	94.12	95.00	94.12
NPV (%)	98.70	95.00	97.44	90.48	96.20	86.36	95.00	82.61	95.00	82.61

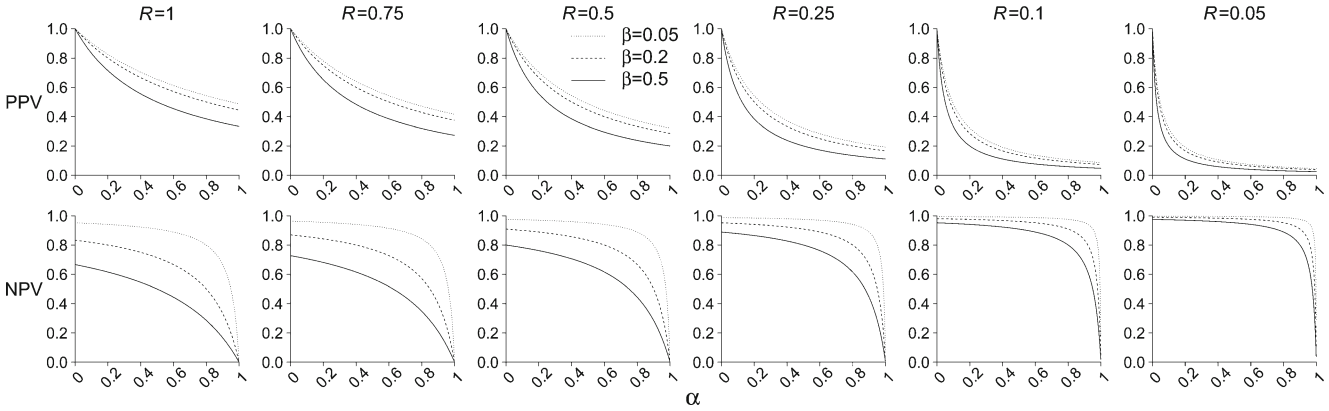

Fig. 2 PPV and NPV as a function of R, α as well as β

in the present study but one might wish to consider this during other investigations.

Finally, the PPV and NPV were calculated (see equations below) for each treatment separately for the experimental design developed based on a priori knowledge and power analysis (i.e., adapted sample size), and also for a test design assumed to follow the OECD guideline 211 strictly (Table 1). The PPV can be calculated as follows (Newman 2008):

$$\mathrm{PPV} = \frac{(1-\beta)R}{R-\beta R+\alpha} \tag{4}$$

The a priori NPV can be derived from Ioannidis (2005):

$$\mathrm{NPV} = \frac{1-\alpha}{1+\beta R-\alpha} \tag{5}$$

However, R can be very low in some research fields (Ioannidis 2005) resulting in low PPV and high NPV (Fig. 2). The β influences the NPV more substantially than the PPV in the conventional NHST design (Fig. 2). This underscores the importance of a priori power analysis to define the minimal adequate sample size per treatment that ensures α and β error rates at a critical ES. All of these have to be pre-defined based on defensible judgment and the considerations outlined previously.

Results

The definitive experiment also followed the OECD protocol except the sample size was dictated by the power analysis. The results again revealed at a concentration of ≥400 μg/L lindane statistically significant negative effects compared to the control with an ES of approximately 15 % (Fig. 3), suggesting a probability of getting these or more extreme data given the H_0 is true is below 5 %. However, most researchers are interested in estimating the probability of H_0 of no adverse effects given the results. If the PPV is considered (Table 1), it can be concluded that the probability of the statistically significant finding, with an ES of 25 % or higher, being true at a particular concentration is 95 %. As an ES of 25 % was found at 500 μg/L, not at 400 μg/L, the PPV was 95 % only at the latter lindane treatment, while a statistically, but not biologically, significant effect was obtained for the 400 μg/L treatment. From the opposite vantage, there is only a 5 % chance that although the statistical test outcome indicates significance together with an ES of 25 %, the true ES is less than 25 % and as defined a priori, not demographically critical. Similarly, performing the *D. magna* reproduction assay according to the OECD—in terms of sample size and thus uncontrolled β, the probability for falsely obtaining a positive (i.e., statistically significant) test outcome

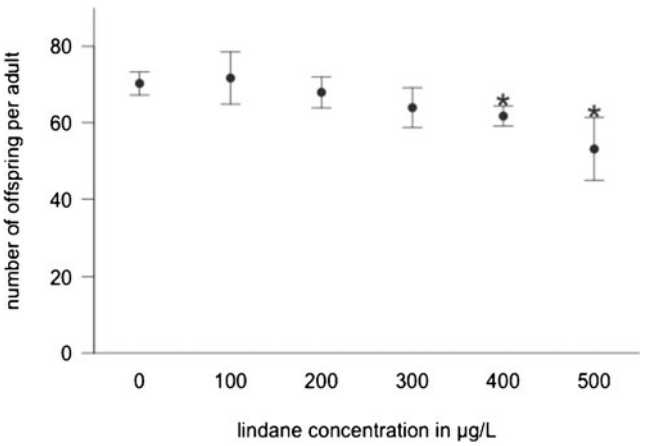

Fig. 3 Mean (±95 % confidence intervals; n=32/14) number of offspring released per adult *D. magna* at different concentrations of lindane during the definitive experiment. *Asterisks* (*) denote statistically significant differences (experimentwise α=0.05) from the control based on a Dunnett's multiple comparisons test

would increase only slightly to 5.9 % (PPV=94.1). In contrast to the PPV, the NPV at the 400 μg/L treatment was only 82.6 % if the test design followed OECD guideline 211 strictly. This corresponds to a probability of approximately 17.4 % of obtaining a statistically non-significant test outcome although the ES truly exceeds the respective threshold. Controlling β by adjusting sample sizes, the NPV increased to 95 % at 400 μg/L lindane, meaningfully elevating the probability that the true ES is actually below the threshold of 25 %, given a statistically non-significant test outcome. Hence, the risk of judging a concentration of a chemical substance as "biologically and environmentally safe" when it is actually unsafe is more than threefold higher for the OECD protocol compared to the test design adapted for the definitive experiment (Table 1).

Discussion

Because both the PPV as well as the NPV estimate the probabilities that a biologically meaningful and a priori defined ES is exceeded given statistically significant and statistically non-significant test outcomes, both might be suitable and feasible extensions of the frequently used statistical methods of NHST. This recommendation is based on controlling of both α and β by power analysis, but also the use of a priori knowledge (R). Moreover, they provide information about the reliability of the experimental data akin to that pervasive in the current human health and clinical science literature (e.g., Altman and Bland 1994). In the present example however, the PPV did not increase substantially if the sample size is determined by an a priori power analysis. This could be a consequence of the fixed α (0.05) and the relative insensitivity of the PPV to values of β and R used in the calculations (Fig. 2). Nevertheless, if β is, for instance, as low as 0.5, which may be the case for other standardized reproduction tests (van der Hoeven 1998), the PPV decreases to approximately 90 % if R is 1 and to 83 % if R is 0.5. This indicates that the PPV for other standardized laboratory experimental designs, which are less powerful than the *D. magna* reproduction assay, may be very low due to neglecting power analysis and not designing tests to reduce β. In contrast to the PPV, the NPV was improved substantially during the present study, increasing the inferential reliability of statistically non-significant test outcomes (Fig. 2). By setting β at 0.05 instead of 0.20 (or by increasing α), as in the present study (Table 1), the NPV increased to 95.0 % at the 400 μg/L treatment (and even higher values for the lower lindane concentrations). Although the necessary experimental efforts increased from a total sample size of 60 to 102, the interpretation of the experimental results is enhanced fundamentally, allowing predictions of "environmentally safe" concentrations of chemicals with a known level of uncertainty (=reliability).

Nevertheless, it needs to be mentioned that obtaining R may be problematic if no information exists regarding the endpoint of interest. We suggested above that it might be suitable to consider information obtained during studies with related species and endpoints for the calculation of R. However, this would considerably increase the uncertainty associated with R. Additionally, one may claim that publication bias – usually against publication of statistically non-significant results – hampers accurate estimation of R (Ioannidis 2005; Newman 2008). Hence, Ioannidis (2005) suggested a correction factor as this uncertainty should eventually be reflected in the PPV and NPV allowing for an assessment of a substance with a known level of (un)certainty. In situations with absolutely no prior knowledge, e.g., a first experiment with a newly developed plant protection product, obtaining a scientifically defensible R represents an important challenge. Under these circumstances, it may be suggested to calculate the PPV and NPV for the latter experiments required for authorization and hence not during the initial experimental stages. A further pitfall regarding R can occur in situations where only a few studies report adverse effects, while a vast number of published data do indicate no effects at a certain concentration of a compound, which might be expected for non-standardized laboratory toxicity tests. This would decrease the PPV and at the same time increase the NPV and could result in further misinterpretation and also misuse. How this can be considered (e.g., by allocating a weight to each study) in the calculation of R is beyond the scope of the present study and needs to be discussed further.

Although widely acknowledged, shortcomings of conventional NHST have led to suggestions that it is not to be used [including utilization to derive NOEC/LOEC (Landis and Chapman 2011)], these methods will not disappear soon from the scientific literature. Authors who continue to utilize NHST are urged to report PPV and NPV if R can be estimated, but minimally the a priori statistical power (or β) together with an ES deemed to be biologically important to detect. This will allow other researchers to make their own judgment on the effects reported as well as the reliability of the experimental data. This additional information would increase transparency about (un)certainty in the data, finally facilitating decisions during environmental risk assessment of chemicals.

Acknowledgments The authors acknowledge S. Hartmann, T. Schell, and S. Schneider for actually performing the experiments as well as the Fix-Stiftung Landau for the financial support regarding the research infrastructure. J.P. Zubrod received funding through a scholarship of the German Federal Environmental Foundation (Deutsche Bundesstiftung Umwelt).

References

Altman DG, Bland JM (1994) Diagnostic tests 2: predictive values. British Med J 309:102

Altman DG, Machin D, Bryant TN, Gardner MJ (2000) Statistics with confidence, 2nd edn. BMJ Books, Bristol

Antunes SC, Castro BB, Goncalves F (2004) Effect of food level on the acute and chronic responses of daphnids to lindane. Environ Pollut 127:367–375

ASTM (2007) ASTM E729-96: Standard guide for conducting acute toxicity tests on test materials with fishes, macroinvertebrates, and amphibians. ASTM International, West Conshohocken, PA, 2007, doi:10.1520/E0729-96R07

Bundschuh M, Zubrod JP, Seitz F, Newman MC, Schulz R (2011) Mercury-contaminated sediments affect amphipod feeding. Arch Environ Contam Toxicol 60:437–443

Crane M, Newman MC (2000) What level of effect is a no observed effect? Environ Toxicol Chem 19:516–519

DeCoen WM, Janssen CR (1997) The use of biomarkers in *Daphnia magna* toxicity testing. IV. cellular energy allocation: a new methodology to assess the energy budget of toxicant-stressed Daphnia populations. J Aquat Ecosyst Stress Recovery 6:43–55

Dunnett CW (1955) A multiple comparison procedure for comparing several treatments with a control. J Am Stat Assoc 50:1096–1121

Ferrando MD, Sancho E, Andreumoliner E (1995) Effects of lindane on *Daphnia magna* during chronic exposure. J Environ Sci Health Part B 30:815–825

Fox DR (2009) Is the ECx a legitimate surrogate for a NOEC? Integr Environ Assess Manag 5:351–353

Fox DR (2010) A Bayesian approach for determining the no effect concentration and hazardous concentration in ecotoxicology. Ecotoxicol Environ Saf 73:123–131

Gigerenzer G (2004) Mindless statistics. J Socio-Econom 33:587–606

Ioannidis JPA (2005) Why most published research findings are false. Plos Med 2:696–701

Jager T (2012) Bad habits die hard: the NOEC's persistence reflects poorly on ecotoxicology. Environ Toxicol Chem 31:228–229

Kline RB (2004) Beyond significance testing: reforming data analysis methods in behavioral research. American Psychological Association, Washington

Landis WG, Chapman PM (2011). Well past time to stop using NOELs and LOELs. Integr Environ Assess Manag 7:vi-viii

Mudge JF, Baker LF, Edge CB, Houlahan JE (2012) Setting an optimal alpha that minimizes errors in null hypothesis significance tests. PlosOne 7:e32734

Nakagawa S, Cuthill IC (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev 82:591–605

Nakagawa S, Forster TM (2004) The case against retrospective statisical power analyses with an introduction to power analysis. Acta etholog 7:103–108

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

Newman MC (2013) Quantitative ecotoxicology. CRC/Taylor & Francis, Boca Raton

OECD (2008) OECD 211: *Daphnia magna* reproduction test. OECD Publishing, Paris

Preuss TG, Hammers-Wirtz M, Ratte HT (2010) The potential of individual based population models to extrapolate effects measured at standardized test conditions to relevant environmental conditions-an example for 3,4-dichloroaniline on *Daphnia magna*. J Environ Monit 12:2070–2079

van der Hoeven N (1998) Power analysis for the NOEC: what is the probability of detecting small toxic effects on three different species using the appropriate standardized test protocols? Ecotoxicology 7:355–361

Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N (2004) Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Nation Cancer Inst 96:434–442

Zubrod JP, Bundschuh M, Feckler A, Englert D, Schulz R (2011) Ecotoxicological impact of the fungicide tebuconazole on an aquatic decomposer-detritivore system. Environ Toxicol Chem 30:2718–2724