View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by UCL Discovery

# Statistical Power Considerations Show the Endocrine Disruptor Low-Dose Issue in a New Light

*Martin Scholze and Andreas Kortenkamp*

The School of Pharmacy, University of London, London, United Kingdom

**Background:** The endocrine disruptor field has been vexed by difficulties in reproducing various claims of effects at unusually low doses. In previous analyses, variations in control responses from experiment to experiment and problems with observing effects in positive controls have been identified as possible explanations of the resulting impasse.

**Objective:** In this article, we argue that both of these viewpoints fail to take sufficient account of the problems that exist in estimating low effects and low-effect doses. We have carried out post hoc power analyses on selected published data to illustrate that claims of low-dose effects (or their absence) are often compromised by insufficient statistical power of the chosen experimental design.

**Conclusions:** We demonstrate that low-dose estimates such as the no observed adverse effect levels derived from statistical hypothesis-testing procedures are dependent on the specific experimental conditions used for testing. Thus, below the statistical detection limit of the experiment, the presence of effects can neither be proven nor ruled out. Common practice is to attempt to establish "doses without effect." However, low-dose estimations in the endocrine-disruptor field could be improved if decisions regarding the toxicologic effect size of relevance formed the starting point of testing procedures. Statistical power considerations could then reveal the resources necessary to demonstrate effect magnitudes of concern.

**Key words:** benchmark, endocrine disruptors, LOEL, low-dose, NOEL, regression, threshold. *Environ Health Perspect* 115(suppl 1):84–90 (2007). doi:10.1289/ehp.9364 available via *http://dx.doi.org/* [Online 8 June 2007]

Difficulties in reproducing experimental observations not only between different laboratories but also within the same laboratory have vexed the endocrine-disruptor (ED) field in recent years. This has provoked an unusually heated controversy in the field, with claims of bias due to sources of research funding (vom Saal and Hughes 2005). In a reanalysis of the published data, Ashby et al. (2004) identified variations in control responses as a major factor explaining the lack of experimental reproducibility. Examples were presented where the between-experiment variability of untreated controls exceeded that seen in treatment groups. At the same time, the effects observed in exposed groups were often consistent from experiment to experiment, even between different laboratories. What is perplexing to many observers is that some laboratories apparently failed to confirm endocrine effects, even when purportedly the same experimental protocol was used. To resolve these problems, Ashby et al. (2004) suggested researching the reasons behind control variability and proposed to set up a historical database that could serve to compare assay performance and test results. In contrast, vom Saal and Welshons (2006) and vom Saal and Hughes (2005) dismissed these suggestions by pointing to a multitude of factors that may influence the magnitude of a response, including control responses. They maintained that background variability in control responses does not, in itself, invalidate experimental observations, as long as the experimental system remains sensitive to positive control agents used for the effect in question. To complicate matters further, there are many examples in which researchers attempted to replicate findings by others, without paying attention to sources of systematic error such as different animal strains or feeds [for some recent examples, see Oehlmann et al. (2006), Ohsako and Tohyama (2005), and vom Saal and Hughes (2005)].

Although both of these viewpoints have certain merits, they do not take sufficient account of issues of statistical power in the testing of ED chemicals (EDCs), nor do they adequately confront the general problems that exist in estimating low effects and low-effect doses. In this article, we wish to place the EDC low-dose discussion in the wider context of quantitative approaches for low-effect estimation, which are used in human toxicology and environmental ecotoxicology, and we will discuss their inherent problems. We argue that satisfactory solutions to the low-dose issue will have to go further than analyzing positive or negative controls, as suggested by Ashby et al. (2004) and vom Saal and Welshons (2006). Claims of absence of effects should include appropriate statistical power calculations, and prospective power evaluations embedded in the design stage of the experiment (or of replication studies) are to be encouraged. Such power evaluations would trigger a badly needed discussion about the choice of effect levels that are judged to be of toxicologic relevance in the context of endocrine disruption.

## Materials and Methods

***Concepts and terms in statistical hypothesis testing.*** Statistical hypothesis testing operates on the basis of the general scientific approach of disproving unsatisfactory hypotheses and proposing new, improved hypotheses that must always be testable. The topic is usually introduced within the framework of the Neyman-Pearson approach, where it is proposed that one of two hypotheses, the null hypothesis and the alternative hypothesis, must hold. The null hypothesis is assumed to be correct, and the goal of a statistical test is to reject the null hypothesis in favor of the alternative hypothesis. Applied to EDC testing, a typical null hypothesis results when—under the conditions tested—the response to the putative EDC is equal to the background response in unexposed control animals (or cells). In statistical terminology, the aim is to reject the notion that "all mean responses are equal" by comparing the mean response of the control group with that of one or more treatment groups. It is important to realize that all test decisions refer to means, and all generalizations from these sample-based results to the population level are only justified in terms of means.

If the null hypothesis is rejected because a chemical shows ED effects, when in fact it does not, a so-called type I error, or false positive, has occurred. Scientists control the probability of type I errors by choosing an appropriate significance level (α, by convention at least 0.05). Much less attention is paid to the probability (β) of committing a type II error (false negative), which is equivalent to saying "chemical X has no ED potential," when in truth it does. Power is the complement of the type II error

This article is part of the monograph "Endocrine Disruptors—Exposure Assessment, Novel End Points, and Low-Dose and Mixture Effects."

Address correspondence to M. Scholze, The School of Pharmacy, University of London, 29-39 Brunswick Square, London WC1N 1AX, United Kingdom. Telephone: 44-20-77535908. Fax: 44-(0)20 7753-5811. E-mail: martin.scholze@pharmacy.ac.uk

This work is part of the European Union-supported EDEN-project "Endocrine Disrupters: Exploring Novel Endpoints, Exposure, Low Dose- and Mixture-Effects in Humans, Aquatic Wildlife and Laboratory Animals" (QLK4-CT-2002-00603) and financial support from the European Commission is gratefully acknowledged.

The authors declare they have no competing financial interests.

Received 22 May 2006; accepted 26 September 2006.

rate (1 – $\beta$) and can be defined as the probability that the experiment will detect a real difference between exposed subjects and controls. In toxicology and risk assessment, the two most well-known descriptors derived from hypothesis testing are the lowest observed (adverse) effect level [LO(A)EL] and the no observed (adverse) effect level [NO(A)EL]. The LO(A)EL is the lowest tested dose showing effects significantly different from untreated controls, and the NO(A)EL is the next lower tested dose. Therefore, decisions about a NO(A)EL depend heavily on the spacing of doses below the LO(A)EL. If the lowest tested dose already causes effects, a NOAEL, in the strict sense of the word, cannot be established experimentally. In such a situation, investigators may wish to retest lower doses. Alternatively, a NOAEL can be estimated by applying extrapolation factors (e.g., as a dose 10 times below the LOAEL).

A key outcome of the ED low-dose peer review under the auspices of the National Toxicology Program (NTP 2001), was that some EDCs exhibit nonmonotonic dose–response relationships, where responses increased as the doses were lowered, resulting in U-shaped dose–response curves. It should be noted that monotonicity is not an *a priori* prerequisite for statistical testing methods.

***Post hoc power analysis.*** We used data from several low-dose studies performed by Sharpe et al. (1995, 1998) and Ashby et al. (1997, 2004) for statistical post hoc power analyses. For power calculations, we used the statistical methods employed by Sharpe et al. and Ashby et al. in their articles. Exact power computations for this test were performed as described by O'Brien and Muller (1993). It should be noted that in their studies, Sharpe et al. (1995, 1998) and Ashby et al. (1997, 2004) did not adjust the individual error rates $\alpha$ for multiplicity, nor did they consider litter effects in the data analysis. In the present study, we also used the dose–response data set for effects of nonylphenol (NP) on vitellogenin (VTG) induction in rainbow trout (Thorpe et al. 2001) for reanalysis with two different multiple hypothesis-testing methods, the Dunnett test (Dunnett 1955) and the Bartholomew test (Bretz and Hothorn 2003).

## Results and Discussion

***Reproducibility, biological variation, and statistical power.*** The problems reported in reproducing some EDC effects might suggest that there is something inherently "difficult" about ED end points that renders them fragile to independent confirmation. With the EDC doses frequently investigated, many of the relevant end points have a narrow dynamic range, and it is not uncommon that the variation in responses found at a given dose level already covers a large part of the maximally possible effect difference between controls and treated groups [see Ashby et al. (2004) and vom Saal and Welshons (2006) for examples]. These features are likely to impact negatively on the probability of detecting an effect if it is present, and may contribute to difficulties in reproducing observations. This would also suggest that there is nothing inherently specific about ED end points and their reproducibility, rather that the small magnitude of effects is a limiting factor. However, this conjecture remains to be examined.

In toxicology, with its focus on safety assessments, reports of the absence of effects pose a difficult dilemma: either the effect truly did not occur, or the chosen experimental system was inadequate to detect any responses. Statistical power considerations are useful in aiding rational decision making in such situations, but power analyses have rarely been applied to the analysis of ED data.

Statistical power, defined as the likelihood of detecting an effect if it is present, is influenced by the sample size, the variance of the effect studied, the difference between the means of the two treatment groups, and the type I error rate ($\alpha$). Power usually increases with sample size, and is higher when the effect differences between treatment groups are large. It decreases with high variance, small differences between treatment groups, and small type I error rates. The rates for type I errors ($\alpha$) and type II errors ($\beta$) are inversely related: the smaller the probability of one, the larger the likelihood of the other. This latter point is of particular relevance to the ED field: Researchers tend to control type I error rates by adopting a small $\alpha$, without realizing that this may have a detrimental effect on power by resulting in an attendant increase in the type II error rate $\beta$.

***Post hoc power analyses of selected EDC studies.*** To illustrate this point, an often quoted example (Ashby et al. 2004) for problems with reproducibility of ED effects is a study of the effects of gestational exposure of rats to octylphenol (OP) and butyl benzyl phthalate (BBP) in which small but repeatable decreases of testis weight and sperm production in Wistar rats were ascribed to these chemicals (Sharpe et al. 1995). Motivated by a failure to observe any of these effects with BBP (Ashby et al. 1997), Sharpe et al. (1998) communicated their experiences with a temporal decline in body and testis weights of control rats. These unexplained changes (which coincided with a change in water supply) were of a magnitude comparable with the most marked treatment effect in the original study (Sharpe et al. 1995).

Table 1 shows a compilation of some of the relevant observations (Sharpe et al. 1995, 1998). In the original study (Sharpe et al. 1995), all treatment effects reached statistical significance (two-sided *t*-test, $\alpha < 0.05$), and the statistical power was sufficient to detect changes in testes weights as a consequence of gestational exposure to OP and BBP. During the period when testes weights in unexposed control animals were low, Sharpe et al. (1998) carried out a repeat experiment with OP. Not only did OP fail to induce a reduction in testis weight, it paradoxically caused the opposite effect by increasing testes weights by 7% (Table 1). Judging by the customary significance criterion of $\alpha = 0.05$, the effect was not statistically significant. However, because fewer offspring were used than in the original studies (Sharpe et al. 1995), the statistical power of this repeat experiment was too low (0.61) to detect an effect with reliability, if it was there. A power of 0.6 is equivalent to finding an effect 6 of 10 times, but missing it 4 of 10 times. Thus, the experimental design used in the repeat study came dangerously close to a 50% chance of reaching the correct decision, if the null hypothesis (OP has no effect on testis weights) was false. Unfortunately, it would have been more efficient, but just as accurate, to flip a coin to make decisions about ED effects of OP, rather than carrying out the experiment. Had the same number of controls and treated offspring been used in the original study, power would have decreased from 0.83 to 0.34, with the likely consequence that Sharpe et al. (1995) would have overlooked the reported OP effects altogether. Figure 1 shows how power increases as mean testis weights decrease for three of Sharpe et al's experiments. This plot allows us to determine what decrease in testis weight would be detectable with various differing degrees of certainty. Using a

**Table 1.** Post hoc power analysis of the effects of OP, BBP, and DES on testes weights in rats.

| Treatment | No. | Absolute testis weight (mg ± SD) | Power[a] |
|---|---|---|---|
| Control | 26 | 2,014 ± 155 | |
| OP (1 mg/L) | 27 | 1,899 ± 123 | 0.833 |
| BBP (1 mg/L) | 35 | 1,809 ± 126 | 1 |
| DES (0.1 mg/L) | 26 | 1,750 ± 180 | 1 |
| Repeat study during the period of low control weights | | | |
| Control | 7 | 1,824 ± 79 | |
| OP (1 mg/L) | 15 | 1,950 ± 173 | 0.61 |
| Repeat studies after the change in control weights normalized | | | |
| Control | 12 | 2,050 ± 84 | |
| DES (0.05 mg/L) | 10 | 1,903 ± 146 | 0.745 |

Adapted from Sharpe et al. (1995, 1998).

[a] Assuming normally distributed testes weights, $\alpha = 0.05$ and a two-tailed *t*-test.

power of 0.8 as the statistical decision criterion, it was possible to detect reduced testis weights of at least 1,904 mg for 1 mg/L OP, 1,908 mg for 1 mg/L BBP and 1,882 mg for 0.1 mg/L diethylstilbestrol (DES) as statistically significant.

After control organ weights had normalized, Sharpe et al. (1998) carried out a repeat study with DES. Probably because of the lower dose of 50 μg/L in drinking water, the reduction in absolute testis weight was smaller this time, albeit still statistically significant. The power of this experiment (0.745) was sufficiently large. Although Sharpe et al. seemed to be concerned about between-experiment variations, it can be argued that the outcome of the DES repeat study agreed reasonably well with the original experiment, considering all of the potential sources of experimental error. Thus, one of the reasons for the perceived inconsistencies in experimental outcomes is statistical power and not, as Sharpe et al. (1998) suggested, an inherent difficulty in reproducing ED data in different laboratories or between different studies.

The above analysis fails to explain why Ashby et al. (1997) did not observe an effect of BBP on testis weights of the male offspring of exposed pregnant rats. Because different strain of rats was used (AP rats, not the Wistar strain employed by Sharpe et al.), and the rats were examined at a later stage [postnatal day (PND) 90 rather than PND20], Ashby et al. themselves prefer to class their experiment as a failure to confirm Sharpe's observations (Sharpe et al. 1995), rather than a refutation of Sharpe et al.'s findings. This is prudent, given that the protocol adopted by Ashby et al. (1997) introduced the potential for systematic differences between the two studies, thus violating a precondition of hypothesis testing within the Neyman-Pearson framework, namely, that under the conditions tested all mean responses are equal. Nonetheless, the study by Ashby et al. (1997) was considerably larger than that by Sharpe et al. (1995). The control group included 109 pups, and the BBP-treated group contained 105 pups. At PND90, the control testes weights (mean ± SD) were 1.66 ± 0.13 g, whereas the same measures in the BBP-exposed group were 1.68 ± 0.12 g. Setting α at 0.05 and assuming normally distributed testes weights [as done by Ashby et al. (1997)], the statistical power of this experiment is only 0.213 for detecting an effect difference of 0.02 g as statistically significant. With the reported effect differences between controls and BBP-treated groups, it would have been necessary to use unsustainably large sample sizes (639 controls and 589 BBP treated) to reach a power of 0.8, a value that is often deemed appropriate. Conversely, the sample sizes chosen by Ashby et al. (1997) would have been sufficient to demonstrate a 3% change in testes weights with a power of 0.8.

This example highlights several important issues. First, the demand that a small type I error (α) is a prerequisite for claims that an effect is present should be matched by an equal requirement for small type II errors (β) (and conversely, high power) for declarations of absence of effects. Strategies for balancing type I and type II errors in toxicologic studies have been discussed (Muller et al. 1984; Muller and Benignus 1992). Second, the importance of making distinctions between statistical significance and toxicologic relevance in interpreting the outcome of ED studies becomes pressing. Power analyses can help in reaching decisions about the sensitivity of experimental studies, but this should not be taken exclusively in the sense of varying sample size. Another important determinant of power is the difference in effect size between treatment groups. These differences should be of toxicologic relevance rather than trivially small. For example, before devoting massive resources to examining ever-smaller effects, a decision should be made whether, for example, a 3% decrease in testes weights should be considered trivially small or of relevance for risk assessment.

***Minimum significant differences.*** Another way of using power analyses is to establish effect differences that can be detected with high probability as statistically significant, and to analyze how this varies with changes in sample size and variance. Such effect differences are termed "minimum significant difference" (MSD) and characterize the statistical detection limit of a specific experimental setup. As an example to illustrate problems with variations in control animals, Ashby et al. (2004)

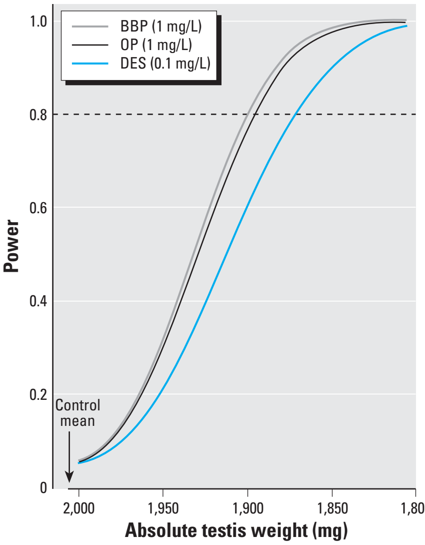

**Figure 1.** Power as a function of observed reduced mean testis weight for three exposures: 1 mg/L OP, 1 mg/L BBP, and 0.1 mg/L DES. Data from Sharpe et al. (1995, 1998).

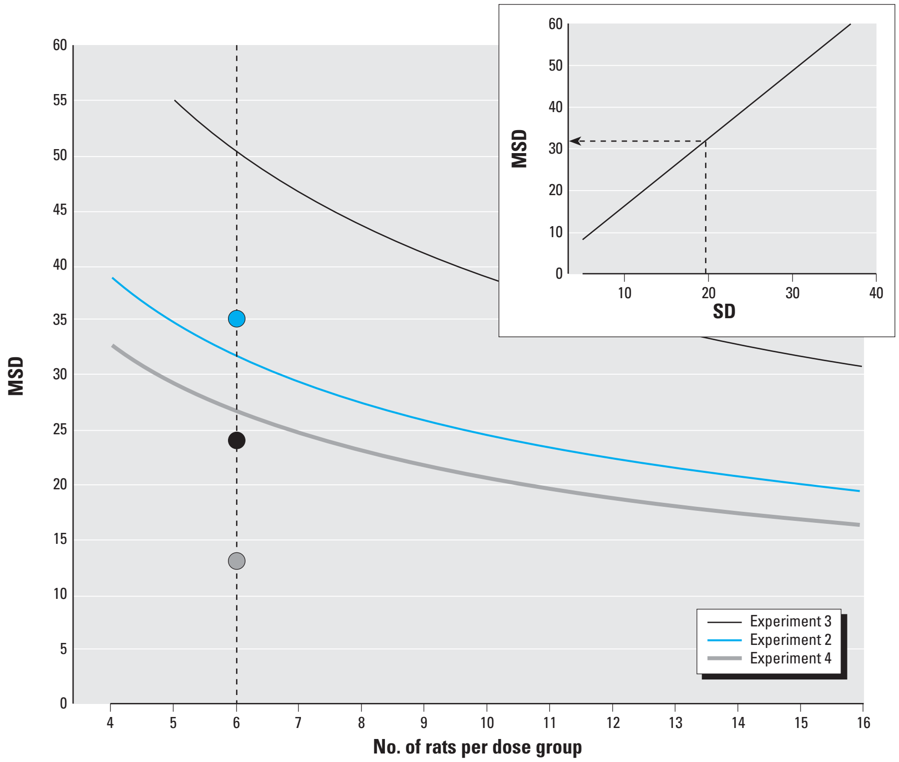

**Figure 2.** MSD for prostate weight as a function of the number of rats per dose group for three 2 μg/kg finasteride treatments carried out in three indpendent experiments reported by Ashby et al. (2004) in their Table 1. Experimentally observed effect differences are shown as data points. The inset shows the MSD as a function of the pooled SD for a balanced design of six animals per group.

reported low-dose experiments of the effects of finasteride in the Hershberger assay; these data present an opportunity to characterize MSDs as a function of the number of animals per dose group.

Finasteride inhibits the conversion of testosterone to dihydrotestosterone and is able to disrupt male sexual development. In combination with testosterone propionate, it also shows antiandrogenic effects in the Hershberger assay; Ashby et al. (2004) aimed to establish a no observed effect level (NOEL) for this end point. They found a statistically significant increase in prostate weights at a dosage of 2 µg/kg finasteride, suggesting an androgenic (and not an antiandrogenic) effect, with an inverted U-shaped dose–response curve. In one repeat study, a statistically nonsignificant increase was observed at this dose, whereas a third experiment failed to show the effect. Pointing to a high variability in control prostate weights, Ashby et al. concluded from these data that the increase in prostate weights observed with 2 µg/kg finasteride was probably a chance finding and that finasteride did not show a low-dose effect because the effect was not reproducible.

For each of the three finasteride low-dose experiments carried out by Ashby et al. (2004), Figure 2 shows how the MSD in prostate weights decreases as the number of animals per dose group increases (two-sided *t*-test, $\alpha < 0.05$). The experimentally observed effect differences are also shown. For the two repeat studies, the observed differences in prostate weight were smaller than the statistical detection limit afforded by the six animals per dose group that were used in these experiments. In the first study, the observed effect was just above the MSD. The inset in Figure 2 shows the dependence of the MSD on effect-data variation, assuming a balanced design with six rats per dose. These considerations support the conclusion drawn by Ashby et al. (2004) that the weight increase observed in the first study was a chance finding.

However, this conclusion needs to be tempered in light of the power analysis shown in Figure 2; within the parameters of the chosen experimental design, the effect was probably a chance finding. With a larger sample size, the observed weight differences would have been resolvable as statistically significant if they were real. The power was simply not sufficient to detect smaller weight differences with confidence. Thus, it was this lack of power and not the variation in control prostate weights per se, that has prevented Ashby et al. from resolving conclusively the finasteride low-dose phenomenon. Without a doubt, the high variation in control values had a negative influence on the power of the experiment, but this could have been compensated for by increasing samples size, up to a limit.

***Dose–response data, sample size, statistical test methods, and NOELs.*** With given effect variance and specified type I and II error rates, the MSD decreases as the sample size increases, with an attendant increase in the sensitivity of the entire experimental set up (Figure 2). However, in practice the MSD never reaches zero. This implies that even with very large sample sizes there will always be an effect difference larger than zero between treated groups and controls that cannot be resolved as statistically significant. This insight has important implications for recognizing the limitations of hypothesis testing in toxicology and of NO(A)ELs, estimates derived from hypothesis testing and one of the most widely used measures of low effects in the ED field and in toxicology.

As previously defined, the NO(A)EL is derived from a LO(A)EL, which is the lowest tested dose that shows effects significantly different from untreated controls, and the NO(A)EL is simply the next lower tested dose. Probably because of its suggestive wording, the term NO(A)EL [described by Moore and Caux (1997) as "one of the most misunderstood notions in ecotoxicology"] is usually taken to imply an absence of effects, as was recently succinctly expressed by Ashby et al. (2004): "If the statistical methods used are appropriate, the absence of significance should indicate the absence of a chemically induced effect."

We used the dose–response data on the effects of NP on VTG induction in rainbow trout by Thorpe et al. (2001) to analyze how estimates of no observed effect concentrations (NOECs) can be influenced by sample size, the statistical method employed for carrying out the significance test, and the chosen significance level $\alpha$ (Figure 3). The original experiment was carried out with 12 fish per treatment group. The mean level of VTG in unexposed fish was 412 ng/mL blood serum. With $\alpha$ and $\beta$ set at 0.05 and 0.1, respectively, and by using Dunnett's test (one-sided), the concentration of NP that yielded an effect statistically significantly different from controls (LOEC; lowest observed effect concentration) can be estimated as 10.2 µg/L. Consequently, the next lower tested concentration, which in this case was 6.09 µg/L, is designated as the NOEC (Figure 3A). Under these experimental conditions, the MSD is 2,510 ng VTG/mL blood serum, which is equivalent to saying that any VTG level < 2,510 ng/mL could not be statistically significant.

Figure 3B represents a hypothetical case in which we examined the influence of sample size on NOEC estimates by simply doubling the original control data to yield 24 untreated fish instead of 12. In this situation, the statistical power would have been sufficient to detect an NP concentration of 6.09 µg/L as statistically significantly different from controls. This concentration was the NOEC estimated from the original data, but the "new" NOEC has now decreased to the next lower tested concentration, 3.57 µg/L. Conversely, the MSD

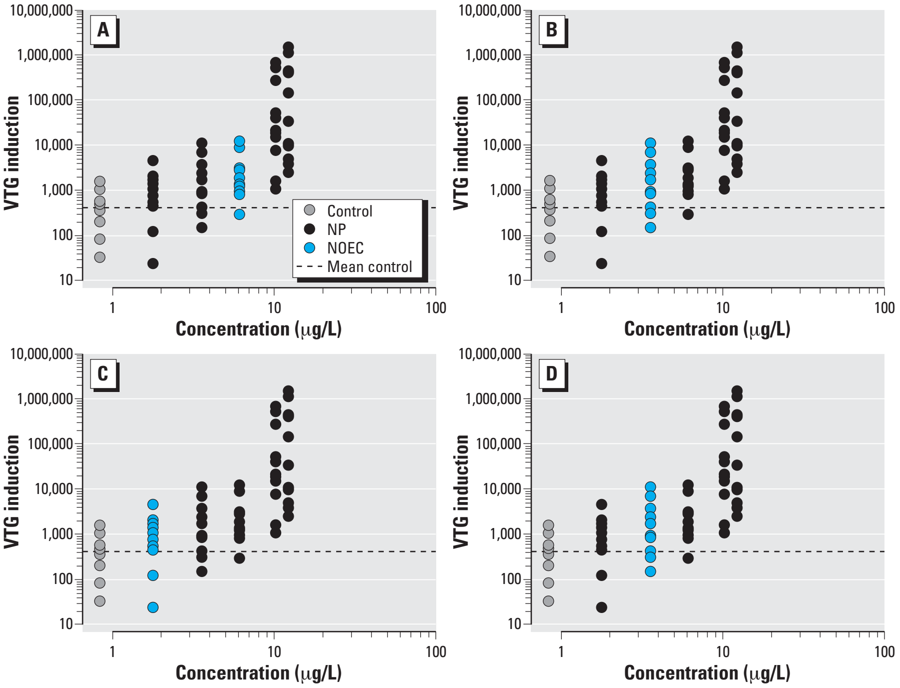

**Figure 3.** NOEC estimations for the dose–response data of the effects of NP on VTG induction in rainbow trout (Thorpe et al. 2001). (*A*) The multiple Dunnett test (one-sided, $\alpha$ = 0.05). (*B*) Same method as in (*A*), but with twice the number of controls. (*C*) Bartholomew test (one-sided, $\alpha$ = 0.05). (*D*) One-sided multiple Dunnett test, but with $\alpha$ = 0.07.

has decreased from 2,510 ng VTG/mL blood serum to 1,260 ng/mL.

The choice of a methodology for testing statistical significance also has a strong effect on what is estimated as the LOEC. There are many statistical procedures for these purposes, and they differ in their assumptions and objectives and in the way they protect against type I and type II errors. For example, the frequently used Dunnett's test is applicable to normally (or log-normally) distributed data without making any assumptions about the underlying dose–response relationship. It controls experiment-wide error rates by correcting for multiple comparisons and is therefore more appropriate than multiple applications of the *t*-test, which would result in unacceptably high type I error rates (false positives). More advanced trend and contrast tests also maintain the global error rate but are more powerful in that they include additional assumptions about the true unknown dose–response relationship (Bretz and Hothorn 2003). For VTG induction in fish, a monotonic trend is plausible, at least for nontoxic concentration ranges; therefore, the application of the Bartholomew test (Bretz and Hothorn 2003) to Thorpe et al.'s original data with 12 fish per group (Thorpe et al. 2001) leads to a different LOEC estimate. Instead of 10.2 µg/L, it is now 3.57 µg/L, and the MSD has changed to 892 ng VTG/mL blood serum (Figure 3C).

Finally, a relatively small but arbitrary change of type I error rate $\alpha$ from the conventional 0.05 to 0.07 is sufficient to decrease the original LOEC of 10.2 µg/L to 6.09 µg/L, with an attendant decrease of MSD from 2,510 ng VTG/mL blood serum to 1,514 ng/L (Figure 3D). It is also evident that the spacing of tested concentrations would have affected the numerical value of an NOEC. Had there been more concentrations tested < 10.2 µg/L, a value > 6.09 µg/L would have been designated as the NOEC. Conversely, omission of the 6.09 µg/L treatment group would have lowered the NOEC to 3.57 µg/L.

These examples demonstrate that NOECs [and by implication also NO(A)ELs] are sensitive to the specific features of the chosen experimental design and the choices of statistical methods and significance criteria. They are not fixed values; thus, when there is no statistically significant difference in response between treated groups and controls, it can only be concluded that the magnitude of effect was below the detection limit of the particular experimental arrangement used. In risk communication terms, this is a weak statement.

Taking this analysis to its logical conclusion, hypothesis testing leads to an irresolvable dilemma. Below the detection limit of a specific experimental system, the presence of effects can neither be proven nor ruled out. We suggest that this, rather than bias due to sources of research funding (vom Saal and Welshons 2006), is at the root of the ED "low-dose" impasse.

***Why worry about small effects? The issue of defining effect thresholds for EDCs.*** The fact that empirical data never allow conclusions of zero effects may, in itself, not be problematic. This is because, from a biological point of view, an effect may be irrelevant even though it is not strictly zero (Slob 1999). Biological systems are capable of correcting certain disturbances provoked by exposure to chemicals; however, the challenge lies in establishing a relevant effect size that defines the borderline between effect and no effect in a biological or toxicologic sense, and not a statistical sense. The issue relates back to the thorny problem of delineating adverse effects from harmless ones and is linked to homeostasis and repair, as well as to the question of how different ED end points relate to one another. For example, Sharpe et al. (1995) attributed decreases in testis weight resulting from gestational exposure to OP or BBP to reductions in the number of Sertoli cells. It would aid the definition of a critical effect size in reduced testis weight if it was known how Sertoli cell number relates to testis weight. Slob (1999) rightly lamented the fact that the topic of establishing relevant or critical effect levels is notoriously neglected, although it should be at the core of toxicology. In the ED field, it has not even appeared on the horizon. In any case, hypothesis-testing procedures could be put on a better footing if decisions about a biological or toxicologic effect size of relevance would form the starting point of power analyses. Power considerations could then reveal which resources are necessary to demonstrate such effects.

***Approaches to defining effects of relevance.*** Several approaches exist to defining an effect of relevance quantitatively. In criterion-referenced evaluations, the importance of an effect is judged in relation to a clear biological or clinical criterion. An example relevant to ED would be the process involved in defining a critical sperm count below which fertility experts recommend assisted fertilization. This was based on information about correlations between sperm count and fertilization success in human populations. The criterion used was the point below which fertilization rates began to decrease with lower sperm counts (Joergensen et al. 2006).

Often, straightforward biological criteria are not available, and in these situations effects of relevance are defined by norm-referenced evaluations. This involves establishing a critical effect by considering the variance of the effect parameter in the population under investigation. There are numerous examples that follow this approach. Cutoff points for elevated cholesterol levels, low birth weights, or late onset of walking in children are all derived by determining certain percentiles—often the 95th percentile—of cumulative population frequency distributions of the selected effect variable. In mutagenicity testing, critical mutation frequencies are defined in terms of multiples of SDs of background mutation rates (Venitt and Parry 1984).

The task of deriving rational criteria for the choice of effect sizes of relevance probably represents one of the biggest challenges in toxicology and the health sciences. For end points relating to ED in human and ecotoxicology, this thinking has not even begun and progress in this area is likely to be slow. Until well-founded criteria emerge, effect sizes may have to be defined arbitrarily (e.g., as 5% or 10% effect levels), as is common practice in the estimation of effect doses by using regression-based methods.

***Estimating low-effect doses: regression-based approaches.*** The weaknesses of hypothesis-testing methods in safety assessment have motivated the search for better procedures in estimating low effects and effect doses, and regression model-based approaches are increasingly promoted as viable alternatives. The rationale is to carry out dose–response analyses to construct a model description for effect data. The model is then used to estimate low-effect doses by either interpolation or extrapolation. These doses are called benchmark doses (BMD) [Crump 1995; U.S. Environmental Protection Agency (EPA) 1995], and they are defined in relation to an effect level that is critical for the end point under investigation. If biological or norm-referenced criteria are not available, the critical level is often set arbitrarily as 5% or 10% of a maximal effect. The lower 95% confidence limit of the BMD is usually referred to as the benchmark dose limit (BMDL) and reflects the degree of uncertainty associated with the data. Poor data quality will lead to a lower BMDL. Conversely, better data, with their reduced degree of uncertainty, are "rewarded" with higher BMDLs. This is very different from hypothesis-testing procedures in which poor data quality usually results in higher NO(A)ELs. For instance, for the dose–response data shown in Figure 3, a corresponding regression fit would produce the same BMD estimate in all cases. In Figure 3B, the statistical confidence interval of the BMD would be slightly smaller because more control data are available. Gaylor et al. (1998) has suggested that, for all of these reasons, the BMDL should replace the NO(A)ELs as a basis for establishing acceptable human exposure limits.

***Regression-based approaches: a solution for the low-dose dilemma?*** Whereas the pair-wise comparisons carried out in hypothesis-testing procedures cannot utilize the information available from other dose groups, the strength of regression methods lies in the fact that the statistical power contained in the entirety of experimental observations is accessible for low-effect dose estimations. However, adequate characterizations of dose–response relationships by mathematical modeling have to rely on the use of a sufficient number of dose groups. Concerns that this might lead to an increase in the total number of animals are unfounded if sound design strategies are used (Woutersen et al. 2001). In any case, meaningful application of regression-based techniques in the ED field would require a drastic change in current testing practice. The overwhelming majority of EDC low-dose studies are not usable for regression analysis because often only one dose (and in exceptional cases two) was tested. In most published EDC studies, the minimum data requirements laid down in technical testing guidelines for NO(A)EL estimations are not fulfilled [e.g., Organisation for Economic Co-operation and Development (OECD) guideline 416 (OECD 2001) recommends a minimum of three doses in two-generation reproduction toxicity studies]. A regression analyses usually requires at least four different doses for nonlinear monotonic dose–response relationships and even more for more complex shapes (e.g., nonmonotonicity).

By way of extrapolation, regression modeling also allows predictions about effects in dose ranges that were not actually tested. However, the validity of low-dose estimates derived from modeling techniques is highly dependent on the correct choice of a regression model (Bailer et al. 2005; Moore and Caux 1997; Sand et al. 2002; Scholze et al. 2001). Unfortunately, *a priori* criteria for choosing a general model for low-dose modeling do not exist, and it is usually not an option to use knowledge about the mode of action of a chemical as a selection criterion for a suitable regression model. In most cases, there is no alternative to choosing a best-fitting model from a collection of regression models on the basis of goodness-of-fit criteria (Scholze et al. 2001; Slob 2001).

Some EDCs exhibit nonmonotonic dose–response curves with inverted U shapes; these can arise when there are dose-dependent changes in the underlying mechanisms (Almstrup et al. 2002). In principle, regression modeling can cope with nonmonotonic dose–response patterns, and appropriate parametric models are available (van Ewijk and Hoekstra 1993). They can be characterized as an expansion of monotonic models by including additional model parameters which allow the estimation of U shapes in the lower part of the curve. However, application of such models to effects of EDs requires testing of a larger number of dose levels than is current practice.

Because the mathematical features of most regression models mean that zero effects are approached asymptotically without the regression line ever crossing the dose axis, the estimation of any possible effect, even down to infinitesimally small values, is feasible in principle. However, because of the lack of power, the models themselves cannot give any indications as to when estimates become unreliable. Statisticians have attempted to overcome this problem by including an additional model parameter that allows the estimation of a mathematical dose threshold associated with a zero response (Cox 1987; Hunt and Bowman 2006; Schwartz et al. 1995). However, considerable confusion has arisen because these modeling outcomes are often interpreted as toxicologic thresholds, but not as what they really are, that is, descriptive model parameters with little predictive power, strongly dependent on the selected model and estimation method. Even with the same set of data, widely differing threshold estimates can be obtained (Slob 1999).

Again, this analysis shows that just like hypothesis testing, regression modeling alone cannot resolve the problems associated with assessing effects of a magnitude below the statistical detection limit (power) of the experiment. The chosen critical effect size should be of sufficient magnitude to allow accurate and precise estimations of BMDs.

***Low-dose estimations using multiple comparisons and regression modeling: an integrated approach.*** With the recognition that doses associated with zero effects cannot be determined empirically, the aim of low-dose EDC testing can only be to derive estimates of doses that correspond to a specific effect magnitude. Thus, the starting point of low-dose testing strategies should be a decision about the effect size a low-dose experiment should be able to demonstrate. Ideally, this decision should be based on biological criteria, but in the absence of viable biological criteria, effect magnitudes can be set arbitrarily.

Only after such a choice has been made can the strengths of hypothesis-testing procedures (the ability to test certain doses with a large number of replicates) and those of regression-based approaches (the ability to assess effect trends) be exploited productively for low-dose testing. It would therefore be desirable to develop a framework for an integrated approach that combines the strengths of multiple comparison techniques with those of regression model-based techniques for the analysis of dose–response data. The key elements of such an approach are outlined in Figure 4. The proposed integrated procedure aims *a*) to identify the minimum effective dose that is statistically significant and that produces an effect that is at least of the relevant effect size, and *b*) if reliable, to estimate the corresponding dose for this effect size (BMD). First, a power analysis is performed with the aim of assessing whether the suggested experimental design is of sufficient sensitivity to demonstrate effect sizes of relevance. This can be achieved by comparing the MSD, which is achievable with the chosen experimental design, with the magnitude of the effect of relevance. In general, only data sets should be used where type I and type II errors are controlled and thus where sufficient power is guaranteed. As a guide, α and β can be set at 0.05 and 0.2, respectively. Estimates of MSD can be made on the basis of multiple historical studies and should be used as a quality control tool. This would answer the concerns of Ashby et al. (2004) about control variability.

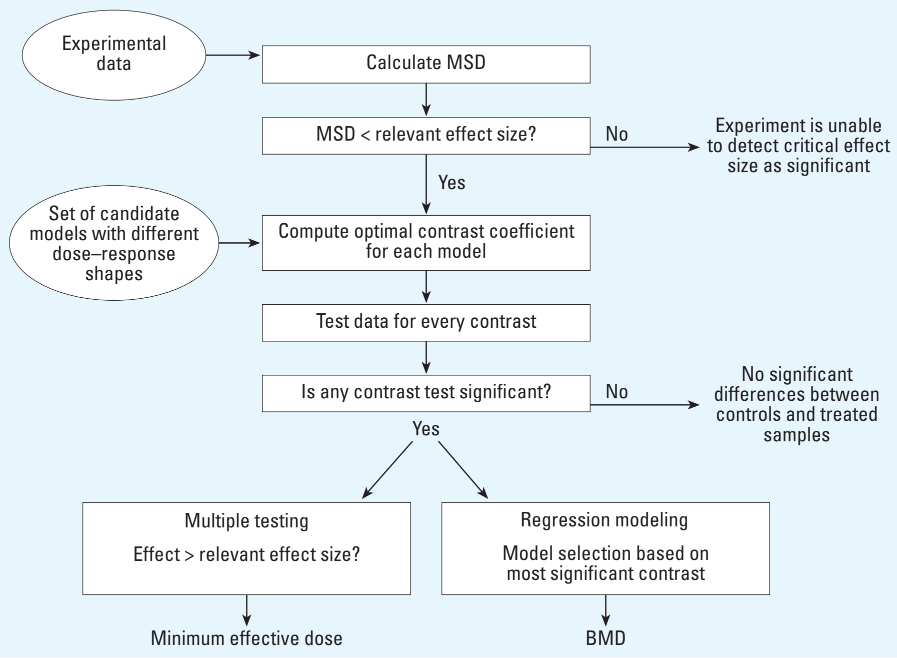
Experimental data
Calculate MSD
MSD < relevant effect size?
No
Experiment is unable to detect critical effect size as significant
Yes
Set of candidate models with different dose–response shapes
Compute optimal contrast coefficient for each model
Test data for every contrast
Is any contrast test significant?
No
No significant differences between controls and treated samples
Yes
Multiple testing
Effect > relevant effect size?
Regression modeling
Model selection based on most significant contrast
Minimum effective dose
BMD

**Figure 4.** Integrated approach for the estimation of effect doses.

Before either hypothesis-testing procedures or regression-modeling techniques are pursued, we suggest so-called multiple contrast tests be carried out (Branson et al. 2003; Bretz et al. 2005) in order to guarantee that at least one tested dose has produced an effect that is significantly different from the controls while preserving the global error rate at the prespecified $\alpha$. These tests make certain assumptions about the underlying shape of a dose–response relationship, which is controlled by the selected contrast (Bretz and Hothorn 2001; Hothorn and Bretz 2000; Neuhaeuser et al. 2000). Because the shape of the underlying dose–response curve is unknown *a priori*, as many contrasts as possible from a pool of *a priori* selected candidate parametric models are selected in order to enable informed choices of the most significant one (which is automatically the most powerful). In this way, the determination of the best contrast for each model is based on the dose regime but not on the effect data. If none of the contrast tests are significant, the null hypothesis cannot be rejected; this indicates that none of the tested doses induced an effect that is statistically significantly different from the MSD or the effect of relevance. Consequently, no further steps toward identifying low-effect doses can be taken.

If, however, the contrast tests signal significance, it is possible to determine the smallest tested dose that produces a response significantly above the relevant effect size. The most significant contrast is the one most likely to represent the underlying dose–response curve and thus can yield a sound criterion for deciding on a suitable candidate regression model for dose–response analyses (Branson et al. 2003; Bretz et al. 2005). At this point, the estimation of low doses can be pursued either by hypothesis testing or by regression analysis. Compared with current practice in the ED field, this procedure has the advantage of transparency and clarity; low-dose estimates are made with clear reference to the statistical power of the experiment.

## Conclusions

There will always be examples in which certain observations cannot be reproduced by other laboratories. As pointed out by Ashby et al. (2004) and vom Saal et al. (2005), some ED effects are specific to the particular strains of animals used for testing, and such specificities are distinct from the failure to reproduce observations when purportedly the same experimental conditions are applied. If steps are taken to ensure that the prerequisite of similarity in experimental conditions is met, we suggest that rigorous statistical power evaluations, fully integrated in the design stage of the experiment, are likely to avoid such situations in the future.

However, there are situations where power considerations are of limited value. Traditionally, all statistical approaches are based on the mean effect concept, which focuses on the middle of the distribution of observed individual responses. It is crucial to consider to what extent this leads to overlooking sensitive subpopulations, which would be found "in the tails" of frequency distributions describing responsiveness to endocrine action.

Nevertheless, power analyses are a valuable tool in recognizing the limitations of specific experimental designs in current low-dose testing. The application of power analyses is likely to promote a badly needed discussion about the magnitude of ED effects that should be considered of toxicologic relevance.

### Reference

Almstrup K, Fernandez MF, Petersen J, Olea N, Skakkebaek NE, Leffers H. 2002. Dual effects of phytoestrogens result in U-shaped dose-response curves. Environ Health Perspect 110:743–748.

Ashby J, Tinwell H, Lefevre PA, Odum J, Paton D, Millward SW, et al. 1997. Normal sexual development of rats exposed to butyl benzyl phthalate from conception to weaning. Regul Toxicol Pharmacol 26:102–118.

Ashby J, Tinwell H, Odum J, Lefevre P. 2004. Natural variability and the influence of concurrent control values on the detection and interpretation of low-dose or weak endocrine toxicities. Environ Health Perspect 112:847–853.

Bailer AJ, Noble RB, Wheeler MW. 2005. Model uncertainty and risk estimation for experimental studies of quantal responses. Risk Anal 25:291–299.

Branson M, Pinheiro J, Bretz F. 2003. Searching for an Adequate Dose: Combining Multiple Comparisons and Modeling Techniques in Dose-Response Studies. Technical Report 2003-08-20. Basel, Switzerland:Novartis Pharmaceuticals. Available: http://www.bioinf.uni-hannover.de/~bretz/paper/TR_MCPMod.pdf [accessed 5 September 2006].

Bretz F, Hothorn LA. 2001. Testing dose-response relationships with a priori unknown, possibly nonmonotone shapes. J Biopharm Stat 11:193–207.

Bretz F, Hothorn LA. 2003. Statistical analysis of monotone or non-monotone dose-response data from in vitro toxicological assays. Altern Lab Anim 31(suppl 1):81–96.

Bretz F, Pinheiro JC, Branson M. 2005. Combining multiple comparisons and modeling techniques in dose-response studies. Biometrics 61:738–748.

Cox C. 1987. Threshold dose-response models in toxicology. Biometrics 43:511–523.

Crump KS. 1995. Calculation of benchmark doses from continuous data. Risk Analysis 15:79–89.

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

Gaylor D, Ryan L, Krewski D, Zhu Y. 1998. Procedures for calculating benchmark doses for health risk assessment. Regul Toxicol Pharmacol 28:150–164.

Hothorn L, Bretz F. 2000. Evaluation of animal carcinogenicity studies: Cochran-Armitage trend test vs. multiple contrast tests. Biom J 42:553–567.

Hunt D, Bowman D. 2006. Modeling developmental data using U-shaped threshold dose-response curves. J Appl Stat 33:35–47.

Joergensen N, Asklund C, Carlsen E, Skakkebek NE. 2006. Coordinated European investigations of semen quality: results from studies of Scandinavian young men is a matter of concern. Int J Androl 29:54–61.

Moore DRJ, Caux P-Y. 1997. Estimating low toxic effects. Environ Toxicol Chem 16:794–801.

Muller KE, Barton CN, Benignus VA. 1984. Recommendations for appropriate statistical practice in toxicological experiments. Neurotoxicology 5:113–125.

Muller KE, Benignus VA. 1992. Increasing scientific power with statistical power. Neurotoxicol Teratol 14:211–219.

Neuhaeuser M, Seidel D, Hothorn L, Urfer W. 2000. Robust trend test with application to toxicology. Environ Ecol Stat 7:43–56.

NTP. 2001. National Toxicology Program's Report of the Endocrine Disruptors Low-Dose Peer Review. Research Triangle Park, NC:National Toxicology Program. Available: http://ntp.niehs.nih.gov/ntp/htdocs/liason/LowDosePeerFinalRpt.pdf [accessed 5 September 2006].

O'Brien R, Muller K. 1993. Unified power analysis for *t*-tests through multivariate hypotheses. In: Applied Analysis of Variance in Behavioural Science (Edwards LD, ed). New York:Marcel Dekker, 297–344.

OECD. 2001. OECD Guideline for Testing of Chemicals. Proposal for Updating Guideline 416 (Two-Generation Reproduction Toxicity Study). Paris:Organisation for Economic Co-operation and Development. Available: http://www.oecd.org/dataoecd/18/13/1948466.pdf [accessed 7 November 2006].

Oehlmann J, Schulte-Oehlmann U, Bachmann J, Oetken M, Lutz I, Kloas W, et al. 2006. Bisphenol A induces superfeminization in the ramshorn snail *Marisa cornuarietis* (Gastropoda: Prosobranchia) at environmentally relevant concentrations. Environ Health Perspect 114(suppl 1):127–133.

Ohsako S, Tohyama C. 2005. Comparison of study controls [Letter]. Environ Health Perspect 113:A582–A583.

Sand S, Filipsson AF, Victorin K. 2002. Evaluation of the benchmark dose method for dichotomous data: model dependence and model selection. Regul Toxicol Pharmacol 36:184–197.

Scholze M, Bödeker W, Faust M, Backhaus T, Altenburger R, Grimme LH. 2001. A general best-fit method for concentration-response curves and the estimation of low-effect concentrations. Environ Toxicol Chem 20:448–457.

Schwartz PF, Gennings C, Chinchilli VM. 1995. Threshold models for combination data from reproductive and developmental experiments. J Am Stat Assoc 90:862–870.

Sharpe RM, Fisher JS, Millar MM, Jobling S, Sumpter JP. 1995. Gestational and lactational exposure of rats to xenoestrogens results in reduced testicular size and sperm production. Environ Health Perspect 103:1136–1143.

Sharpe RM, Turner KJ, Sumpter JP. 1998. Endocrine disruptors and testis development [Letter]. Environ Health Perspect 106:A220–A221.

Slob W. 1999. Thresholds in toxicology and risk assessment. Int J Toxicol 18:259–268.

Slob W. 2001. Dose-response modeling of continuous endpoints. Toxicol Sci 66:298–312.

Thorpe KL, Hutchinson TH, Hetheridge MJ, Scholze M, Sumpter JP, Tyler CR. 2001. Assessing the biological potency of binary mixtures of environmental estrogens using vitellogenin induction in juvenile rainbow trout (*Oncorhynchus mykiss*). Environ Sci Technol 35:2476–2481.

U.S. EPA. 1995. The Use of the Benchmark Dose (BMD) Approach in Health Risk Assessment. EPA/630/R-94/007. Washington, DC:U.S. Environmental Protection Agency.

van Ewijk PH, Hoekstra JA. 1993. Calculation of the EC50 and its confidence interval when subtoxic stimulus is present. Ecotoxicol Environ Saf 25:25–32.

Venitt S, Parry J. 1984. Mutagenicity Testing: A Practical Approach. Oxford/Washington, DC:IRL Press.

vom Saal FS, Hughes C. 2005. An extensive new literature concerning low-dose effects of bisphenol A shows the need for a new risk assessment. Environ Health Perspect 113:926–933.

vom Saal FS, Richter CA, Ruhlen RR, Nagel SC, Timms BG, Welshons WV. 2005. The importance of appropriate controls, animal feed, and animal models in interpreting results from low-dose studies of bisphenol A. Birth Defects Res A Clin Mol Teratol 73:140–145.

vom Saal FS, Welshons WV. 2006. Large effects from small exposures. II. The importance of positive controls in low-dose research on bisphenol A. Environ Res 100:50–76.

Woutersen RA, Jonker D, Stevenson H, Biesebeek JDT, Slob W. 2001. The benchmark approach applied to a 28-day toxicity study with Rhodorsil Silane in rats: the impact of increasing the number of dose groups. Food Chem Toxicol 39:697–707.