Variability in Nontarget Terrestrial Plant Studies Should Inform Endpoint Selection

Jane P Staveley,† John W Green,‡ Josie Nusz,§ Dan Edwards,§ Kevin Henry,¶ Matt Kern,§ Andrew M Deines,† Richard Brain,† Brad Glenn,†† Nathan Ehresman,§§ Tiffany Kung,## Kimberly Ralston-Hooper,### Faith Kee,## and Steve McMaster####

†Exponent, Chapel Hill, North Carolina, USA
‡DuPont, Wilmington, Delaware, USA
§BASF, Research Triangle Park, North Carolina, USA
¶NovaSource/Tessenderlo Kerley, Phoenix, Arizona, USA
#Waterborne Environmental, Leesburg, Virginia, USA
††Syngenta, Greensboro, North Carolina, USA
§§Bayer CropScience, Research Triangle Park, North Carolina, USA
###Metalaxyl Task Force, Morrisville, North Carolina, USA
####FMC Corporation, Ewing, New Jersey, USA
####††Dow AgroSciences, Indianapolis, Indiana, USA
####Industry Task Force II on 2,4-D Research Data, Zionsville, Indiana, USA

ABSTRACT

Inherent variability in nontarget terrestrial plant (NTTP) testing of pesticides creates challenges for using and interpreting these data for risk assessment. Standardized NTTP testing protocols were initially designed to calculate the application rate causing a 25% effect (ER25, used in the United States) or a 50% effect (ER50, used in Europe) for various measures based on the observed dose–response. More recently, the requirement to generate a no-observed-effect rate (NOER), or, in the absence of an NOER, the rate causing a 5% effect (ER05), has raised questions about the inherent variability in, and statistical detectability of, these tests. Statistically significant differences observed between test and control groups may be a product of this inherent variability and may not represent biological relevance. Attempting to derive an ER05 and the associated risk-assessment conclusions drawn from these values can overestimate risk. To address these concerns, we evaluated historical data from approximately 100 seedling emergence and vegetative vigor guideline studies on pesticides to assess the variability of control results across studies for each plant species, examined potential causes for the variation in control results, and defined the minimum percent effect that can be reliably detected. The results indicate that with current test design and implementation, the ER05 cannot be reliably estimated. Integr Environ Assess Manag 2018:14:639–648. © 2018 The Authors. Integrated Environmental Assessment and Management published by Wiley Periodicals, Inc. on behalf of Society of Environmental Toxicology & Chemistry (SETAC)

Keywords: Nontarget plants Testing Variability NOEC ER05

INTRODUCTION

Nontarget terrestrial plant (NTTP) studies facilitate the assessment of potential effects of pesticide active ingredients on vegetation inadvertently exposed to spray drift or runoff. These studies simulate worst-case exposure via high volume overhead spray, either in a preemergent situation, in which soil is sprayed where seeds are planted, or a postemergent situation, in which the foliage of young plants is sprayed. Standardized guidelines for NTTP testing for the preemergent scenario focus on seedling emergence (SE) and growth (i.e., U.S. Environmental Protection Agency [USEPA] OCSPP 850.4100, Seedling Emergence and Seedling Growth [USEPA 2012a]; Organisation for Economic Co-operation and Development [OECD] Test Guideline No. 208 [OECD 2006a], Terrestrial Plant Test: Seedling Emergence and Seedling Growth Test). Vegetative vigor (VV) is the focus of the guidelines considering postemergent exposure (i.e., USEPA 850.4150 [USEPA 2012b], Vegetative Vigor; OECD Test Guideline No. 227 [OECD 2006b], Terrestrial Plant Test: Vegetative Vigour Test). Except for a limit test in which only one exposure rate is examined, the standardized study
guidelines for NTTP testing examine the effects on the test plants of a series of exposures (concentrations of applied spray solutions, i.e., application rates). The resulting dose–response information is used to determine the concentration or rate at which a certain adverse effect in the measured response parameter, relative to the control, is expected to occur in a group of test organisms under specified exposure conditions. In the United States, the test endpoint selected from the dose response is the 25% effect rate, ER25 (for effective rate; sometimes called the EC25, for effective concentration; ER and EC are used interchangeably herein), which is then used in regulatory risk-assessment schemes. Selection of this endpoint stems from the first nontarget plant test guidelines published in 1982 by USEPA’s Office of Pesticide Programs (Holst and Ellwanger 1982), which described 3 testing tiers for assessing the effects of pesticides on nontarget plants. In these guidelines, if Tier I results, which assess the effect of the maximum label application rate, showed growth reduction or visual phytotoxicity of more than 25% (compared to control), then Tier II tests were required. Tier II tests were dose–response tests from which the ER25, ER50, and the no-observed-effect rate (NOER) were derived. According to the 2001 Scientific Advisory Panel Briefing on the Proposal to Update Non-Target Plant Toxicity Testing Under NAFTA (Davy et al. 2001), although a 50% effect level was selected for aquatic plants, which have a shorter recovery period, the ER25 value was thought to be an appropriate endpoint for terrestrial plants, as it “allowed the agency to account for low-level damage of a cosmetic nature to high value ornamentals and fruits.” USEPA recognized that some terrestrial plants may recover from a 25% defoliation or reduction in early growth with no adverse effect on yield while others may not.

The OECD Test Guidelines mention both the ER25 and the ER50. The ER50 is typically used in regulatory risk assessments in Europe. In addition to these regression-based endpoints, both the USEPA and the OECD Test Guidelines mention the NOER (sometimes referred to as the no-observed-effect concentration, NOEC; these terms are used interchangeably herein). This is the highest rate or concentration of a test substance to which organisms are exposed under specified exposure conditions that does not cause a statistically significant adverse effect as compared to the control. The NOER is determined through hypothesis testing. The advantages and disadvantages of the NOEC compared to endpoints that make use of the entire dose response (often termed regression-based endpoints) such as the EC25 and EC50 have been debated extensively (Green et al. 2012; Landis and Chapman 2011; Fox 2011) and are not discussed here; suffice it to say that test designs optimized for regression-based endpoints are not optimal for hypothesis-based endpoints and vice versa (USEPA 2012c; Stephan 1977). Thus, the difficulty of obtaining both types of endpoints from the same study is to be expected.

Nonetheless, the desire to estimate a “no effect” level for use in risk assessment has prompted recent requirements to use the dose–response information from an NTTP test to derive an ER05 or ER10, presumably under the assumption that a 5% or 10% effect level is a more conservative basis on which to regulate than a 25% or 50% effect level. Some of the desire for an ER05 or ER10 endpoint likely also results from the current preference for regression-based endpoints over hypothesis-based endpoints (van Dam et al. 2012; Jager 2012; Landis and Chapman 2011). However, it is important that the selected endpoint is statistically reliable (Green 2015, 2016). Even early work in toxicology (Trevan 1927) identified the concept that the relative error in determination of the ER50 is smaller than ERx for other values of x for many types of assays, a concept that has been expressed recently relative to NTTP studies, where it has been stated that “the uncertainty around an ER50 is smaller than the uncertainty around an ER10” (EFSA 2014). Previous evaluations of variability and statistical detectability in toxicity tests have been conducted for aquatic plant species (Brain et al. 2004, 2005; Hanson et al. 2003; Sanderson et al. 2009; Knauer et al. 2006) and for mesocosms (Kraufvelin 1998) but have not been explored to such a degree for terrestrial plants. Therefore, this investigation was conducted to understand the degree to which an ER05 or ER10 could be “reliably” determined (based on a definition we propose herein), given the inherent variability in control plant growth in these tests. This inherent variability, resulting from the biological attributes of the plants under the test conditions, is composed of both within-treatment variance as well as between-study variance and was statistically explored in this investigation. The issue is particularly critical in the U.S., where, in the absence of a NOER, the USEPA test guidelines state that an ER05 should be determined, and this value would then be used for endangered species assessments (USEPA 2013). Thus, the following questions were posed:

- Is it possible to “reliably” detect a 5% effect (i.e., difference from the controls) that can be attributed to the test substance in standard NTTP studies?
- If not, what level of difference can be detected?
- What is the typical variability in the controls, and does this vary by species and growth parameter?
- What are the potential sources of variability in the controls?

**METHODS**

**Data collection and compilation**

Data were collected from standard guideline SE and VV studies conducted during the period 2004–2015. These studies were conducted under Good Laboratory Practice regulations and in compliance with OECD guidelines 208 and 227 (OEC 2006a, 2006b), respectively, and/or comparable USEPA OCSPP guidelines 850.4100 and 850.4150 (USEPA 2012a, 2012b). For each study, the following information was collected: study sponsor, study identification number, test substance, test facility, year study was initiated, quarter study was initiated, study duration (days), species, application rates...
(value and units), method of weight determination (dry weight or wet weight), number of pots per replicate and number of plants per pot, diameter of pots, adjuvant type (if used) and percent, type of soil, and type of fertilizer (if any). For each replicate of each control and treatment, the following information was collected: mean and standard deviation of plant height at test termination, weight at test termination, percent emergence (SE only), and percent survival.

Nine different crop protection chemical companies provided studies conducted at 13 different laboratories, encompassing 53 pesticidal test substances. Data were compiled from 52 SE and 52 VV study reports, each containing data for multiple species. For any given species and study type, there were 1–39 studies available. For both study types, the predominant growth parameters used to determine study endpoints, dry weight and shoot height, were collated. Some studies reported wet weight instead of dry weight; since wet weight was not used in this analysis, there are an unequal number of results for different growth responses (i.e., more data for shoot height than for dry weight). Table 1 provides the number of studies by species common name, study type, and growth parameter. Survival and emergence data were also compiled; studies that did not meet the validity criteria expressed in the test guidelines (e.g., for control emergence and survival) were not part of the data compilation. Each study included 4–12 replicates, with most studies incorporating 6–8 replicates for the control, adjuvant control (if applicable), and each test treatment. Values that were considered as outliers by the authors of the study reports were omitted from analysis. Where only 1 or 2 studies were available for a given test type and species, these results were removed from further analyses. These included studies with field bean, rice, sorghum, and buckwheat. This resulted in a total of 21 species evaluated for the SE study type and 22 species evaluated for the VV study type. These species are listed in Table 2.

Analysis approach

The primary objective of this paper is to investigate a conditional regulatory requirement to estimate ER05 (or ERx for some other small value of x) for the chosen responses from every study done under the 2 indicated test guidelines. It is thus important to explore a database of such studies that come from many testing facilities under a variety of typical conditions to make sure the conclusions are based on the reality of such testing rather than purely theoretical concerns. To the extent that the database is typical of what is encountered in conducting studies under the indicated test guidelines, the conclusions should inform regulatory authorities as to what they can reasonably require. The database comprised studies conducted on pesticide active ingredients; however, if the same test guidelines were used to examine effects of other substances on the same terrestrial plant species, the conclusions should also be applicable.

The main focus of this analysis was on control values, since there were more control data available than full dose–response data and the variability of control values provides a good indication of the effect size that can be distinguished from random variability (noise) or that may be estimated from regression models. However, there were 45 SE studies and 44 VV studies for which full dose–response data were available and for which it was possible to fit a regression model. The analysis of control data is discussed first, followed by the analysis of the full dose–response data. Finally, a discussion of the analysis of possible explanatory variables (control type, adjuvant used, soil type, fertilizer type, year, and season) is presented.

Variability analysis for control data. After the data were compiled and sorted, the variability in the control response for each combination of test type, plant species, and growth parameter was determined. The test type was either SE or VV, the plant species was one of the species listed in Table 2, and the growth parameter was either shoot height or dry weight. Sorting data by these variables was considered important, since it was suspected that growth might be more variable for some species than others, or the variability within growth parameters may be different, and pooling all the data might obscure these factors. The coefficients of variation (CVs) and minimum detectable difference expressed as a percentage (MDD%) were determined for the controls for each species for each growth parameter (shoot height and dry weight) for both SE and VV tests. We used the MDD% based on control variance as a simple guide to the size effect that can be estimated from a suitable regression model fit to the data from a single test. There are 3 primary applications for use of the MDD%. First, it is useful for planning a study when there is limited data available. For example, when a new chemical is to be tested, control data from previous studies can be used to develop the experimental design. It is also helpful in regulatory review to demonstrate the size effect that can be detected from a given data set when the statistical test employed found no significant treatment-related effect. Finally, it is useful in setting regulatory requirements by demonstrating the size effect that can be reasonably required in an estimate. MDD% has been applied to a class of ecological studies by Brock et al. (2015) and Wieczorek et al. (2017) and used in EFSA guidance (EFSA 2013), USEPA publications (e.g., Harcum and Dressing 2015), and in clinical trials (Meinert 2012). The concept sometimes appears under names such as minimal detectable difference or minimum detectable change. MDD% serves a function analogous to a power calculation. In the latter, given the variability expected or observed in the control, the statistical test to be employed, the experimental design, and the size effect to be detected, a probability of finding a statistically significant effect is calculated. In the same sense, if MDD% = 20, then it is unlikely to be able to detect or estimate an effect of less than 20% and it is likely to be able to detect or estimate an effect of 20% or more. The term “likely” is used here in an informal way and no probability is assigned. The mathematical derivation of the MDD% is presented below, followed by support as to why the MDD% is a reasonable substitute for the minimum
size effect, MSEff, which can be estimated from a regression model subject to the same level of variability within each treatment group as exists in the control.

The CV is a common statistical measure and was determined as shown in Equation 1.

\[
\text{CV} = \frac{s}{\text{Mean}} \times 100
\]  

(Standard Deviation, \(\sigma\)/Mean, \(\mu\) × 100  \(1\))

The MDD% can be regarded as a translation of CV (see Equation 5) into a metric that can be more directly compared to potential effects. MDD% most directly indicates the likelihood of finding a specific size effect statistically significant from a hypothesis test, such as Dunnett, Williams, or a corresponding nonparametric alternative (Dunn or Jonckheere-Terpstra), as demonstrated in the following discussion.

| Species      | VV |     | SE |     |
|--------------|----|-----|----|-----|
|              | Shoot height | Dry weight | Shoot height | Dry weight |
| Barley       | 2  | 2   | 4  | 4   |
| Common bean  | 8  | 8   | 6  | 6   |
| Buckwheat*   | 0  | 0   | 2  | 2   |
| Cabbage      | 18 | 17  | 16 | 16  |
| Carrot       | 10 | 9   | 10 | 8   |
| Corn         | 37 | 34  | 37 | 34  |
| Cucumber     | 20 | 17  | 20 | 17  |
| Field bean*  | 1  | 0   | 1  | 0   |
| Flax         | 3  | 3   | 3  | 3   |
| Lettuce      | 24 | 23  | 22 | 22  |
| Oat          | 21 | 18  | 23 | 20  |
| Onion        | 35 | 32  | 33 | 30  |
| Pea          | 3  | 3   | 4  | 4   |
| Perennial ryegrass | 10 | 10  | 8  | 8   |
| Radish       | 20 | 18  | 18 | 18  |
| Oilseed rape | 25 | 23  | 28 | 25  |
| Rice*        | 1  | 1   | 1  | 1   |
| Rye          | 1  | 1   | 0  | 0   |
| Ryegrass     | 28 | 24  | 30 | 27  |
| Sorghum*     | 0  | 0   | 1  | 1   |
| Soybean      | 37 | 33  | 39 | 36  |
| Sugar beet   | 14 | 11  | 18 | 15  |
| Sunflower    | 5  | 3   | 9  | 7   |
| Tomato       | 35 | 31  | 35 | 32  |
| Turnip       | 3  | 3   | 3  | 3   |
| Wheat        | 18 | 17  | 16 | 16  |

*Species omitted from further analysis due to insufficient data.

A standard formula (e.g., Hogg and Tanis 1988) to compare 2 sample means from normal populations with the same variance and sample sizes is shown in Equation 2.

\[
T = \frac{\text{Diff}}{\sigma \sqrt{\frac{1}{n}}}
\]  

(2)

where \(n\) is the common sample size, Diff is the magnitude difference \((\bar{x}_1 - \bar{x}_2)\) between 2 sample means, \(\sigma\) is the sample standard deviation derived from the pooled estimate of the presumed common variance, and \(T\) is Student’s \(t\)-statistic with df = \(2(n - 1)\) degrees of freedom. Here \((\bar{x}_1)\) is the same as \(\mu\) from Equation 1. Equation 2 can be modified easily to handle the case of unequal variances, but in the present circumstance of basing conclusions on control data, there is little need to do so. In
The percent change from control \( (x_1) \) that can be detected with 95% confidence is 100Diff/\( x_1 \).

Now CV = 100\( \sigma /x_1 \), so \( x_1 = 100 \sigma /CV \), and it follows that the minimum percent change from control that can be detected with 95% confidence is given, more generally, by

\[
\text{MDD\%} = \frac{100 \text{Diff}}{x_1} = \frac{100 \sigma \sqrt{2}}{(100 \sigma /CV) \sqrt{n}} = CV \sqrt{\frac{2}{n}}. \tag{5}
\]

For small sample sizes (e.g., \( n = 6 \)), \( T \) will be somewhat larger than 2, and in a typical nontarget plant study, the 2-sided \( t \)-test will be replaced by the 1-sided Williams, Jonckheere-Terpstra, or Dunnett test. The 1-sided test will tend to reduce the minimum effect size that can be detected, but the lower degrees of freedom and the need to adjust for multiple comparisons of treatments to control rather than just the one used in the \( t \)-test will tend to increase it. Equation 5 is a balance between these offsetting trends and gives a reasonable indication of the actual minimum effect size that can be detected. The article by Brock et al. (2015) contains additional useful discussion.

Equation 5 is easily modified to handle cases where the number of replicates in the control differs from that in treatment groups. Such situations commonly occur when a solvent or adjuvant is used, since in such studies, both a negative control and a solvent or adjuvant control are generally used. According to Green and Wheeler (2013), in these situations, the 2 controls should be pooled for analysis of treatment effects except when there are significant differences between the controls. This practice was followed in the current investigation. The relevant modification in Equation 5 is given by Equation 6:

\[
\text{MDD\%} = \frac{100 \text{Diff}}{x_1} = \frac{100 \sigma \sqrt{2}}{(100 \sigma /CV) \sqrt{n_1} \sqrt{n_2}} = CV \frac{1}{\sqrt{n_1} \sqrt{n_2}}. \tag{6}
\]

For each species, test type, and growth response combination, the distribution of CVs and MDD% was determined, including determination of the mean, median, 25th and 75th quantiles, and minimum and maximum values. To examine the effect of control variability on the value of MDD% that can be "reliably" detected, we selected the 75th quantile of the MDD% and compared this to different effect levels used or suggested for evaluating risk (i.e., ERx, ER15, ER25). This comparison helps to answer the following question: Is the ERx greater than the MDD% in 75% of the studies, such that it is reasonable to use ERx as an endpoint for that combination of test type, species, and growth response? It was considered that a threshold of "75% of the time" provided reasonable confidence in the answer but was not overly restrictive. We thus defined "reliably detected" as the 75th quantile of the MDD% (MDD%75) being less than the stated effect level (i.e., ERx).

Analysis of full dose–response data. Although the MDD% most directly indicates the likelihood of finding a specific size effect statistically significant from a hypothesis test, a relevant question is whether MDD% is also a reasonable estimate of

| Common name | Latin name |
|-------------|------------|
| Monocots    |            |
| Barley      | Hordeum vulgare L. |
| Corn        | Zea mays   |
| Oat         | Avena sativa |
| Onion       | Allium cepa |
| Perennial ryegrass | Lolium perenne |
| Ryegrass    | Lolium multiflorum |
| Wheat       | Triticum aestivum |
| Dicots      |            |
| Common bean | Phaseolus vulgaris |
| Cabbage     | Brassica oleracea |
| Carrot      | Daucus carota |
| Cucumber    | Cucumis sativa |
| Flax        | Linum usitatissimum |
| Lettuce     | Lactuca sativa |
| Pea         | Pisum sativum |
| Radish      | Raphanus sativus |
| Oilseed rape| Brassica napus |
| Soybean     | Glycine max |
| Sugar beet  | Beta vulgaris |
| Sunflower   | Helianthus annuus |
| Tomato      | Lycopersicon esculentum |
| Turnip      | Brassica rapa |

Table 2. Plant species included in variability analysis
the minimum size effect (MSEff) that can be estimated from a regression model subject to the same level of variability within each treatment group as exists in the control. The theoretical derivation of this relationship is discussed in Supplemental Information Attachment 1. Because the full dose–response data for the studies are not included in the Supplemental Information, only an abbreviated treatment of this aspect of the analysis is provided. Briefly, there were 44 VV studies and 45 SE studies for which full dose–response data were available. A suite of nonlinear models consistent with OECD (2014) was used with standard model selection criteria to examine the relationship between MDD% and MSEff, as discussed in Supplemental Information Attachment 1. MDD% values predicted by control variability were compared with regression models fit to full dose–response data by determining the smallest effect size that could be estimated with a confidence interval that does not contain zero from a model judged to be a reliable fit.

Analysis of explanatory attributes. The influence of selected attributes was examined in an attempt to determine whether there were certain study attributes responsible for the variability in growth parameters (e.g., shoot height and dry weight) observed in the controls. These attributes were control type (negative or adjuvant), adjuvant type used, soil type, fertilizer type, year, and season (quarter). There were at least 6 different adjuvants used, plus some studies where the adjuvant was not identified. There were 2 different fertilizers reported to have been used, plus studies where either no fertilizer was used or no information was available. There were 3 soil types identified, plus studies in which the soil type was not identified. Studies were conducted in all 4 seasons (identified as first, second, third, and fourth quarter, with Q1 being January through March, and so forth) and spanned the period 2004–2015, although the distribution over seasons and years was uneven. The soil, fertilizer, and other attributes varied from species to species. Overall, there were too few observations to explore the potential interactions of these explanatory attributes. However, the analysis proceeded as follows.

Two approaches were used to explore the contributions of these potential explanatory attributes to variability. First, a main-effects analysis of variance (ANOVA) was performed separately for each species and type of study (VV or SE) on each response (dry weight and shoot height) with season, soil type, fertilizer, and adjuvant as factors. For those factors that made a significant contribution ($P < 0.05$) to the variance, the mean responses were examined for each level of that factor to determine what differences contributed to the significant factor. The number of studies contributing to each level of the factor was computed. If the factor level was represented by a single study, little confidence was given to the observation.

The second approach was to remove species differences by first standardizing the data for each species to have mean 10 and standard deviation 1. Following that, a main-effects ANOVA was done that ignored species but otherwise had the same 4 factors. Also, a separate ANOVA was done with each of the 4 factors alone. Since standardization is species specific, the relative variability across soil types, fertilizers, etc. is retained. There is no reason to believe that all species will react to the same set of conditions in the same way, but this approach allowed a broader assessment of commonalities that cut across species.

RESULTS AND DISCUSSION

Variability analysis for control data

The data collected were anonymized as to the test substance, study sponsor, and performing laboratory. The raw data for the controls from each study are presented in Supplemental Information Attachment 2.

While the variability in controls differed among species, the range of variability was similar across growth parameters (dry weight and shoot height) and study types (SE and VV) (Table 3, Figures 1–4).

Figures 1–4 present the mean, median, minimum, and maximum values for the MDD% for each species by growth parameter and study type. Figures 1 and 2 show SE results, and Figures 3 and 4 show VV results. In addition to the maximum and minimum, the 25th and 75th quartiles are shown to illustrate where the bulk of the responses lie and to minimize the visual effect of outliers. The value in parentheses after the species name indicates the number of studies in the data set for each species.

Considering the larger MDD% interquantile ranges, dry weight control data tended to demonstrate more variability than did shoot height for both study types, and variability was greater for the SE study than the VV study. While there were differences in the variability among species, clear trends in performance by species were difficult to observe in most instances. Flax displayed relatively high median MDD%. However, due to the relatively low sample size ($n = 3$) for this

| Test type          | Growth response parameter | Range of median CVs (%) | Range of median MDD (%) |
|--------------------|---------------------------|-------------------------|-------------------------|
| Seedling emergence | dry weight                | 9–26.1                  | 13.5–39.5               |
|                    | shoot height              | 5.1–13.4                | 6–21                    |
| Vegetative vigor   | dry weight                | 7.2–23.2                | 9–30                    |
|                    | shoot height              | 3.5–14.5                | 5–14.5                  |
species, strong conclusions cannot be made on this point. No trends in relative control variability were observed for monocot versus dicot species.

To further examine the effect of control variability on the value of MDD% that can be "reliably" detected, the 75th quantile of the MDD% was compared to different effect levels (i.e., ER05, ER10, ER15, and ER25; Table 4). This comparison helps to answer the following question: Is the ERx greater than the MDD% in 75% of the studies, such that it is reasonable to use ERx as an endpoint for that combination of test type, species, and growth response? It was considered that a threshold of "75% of the time" provided reasonable confidence in the answer but was not overly restrictive. As discussed previously, we thus defined "reliably detected" as the 75th quantile of the MDD% (MDD%75) being less than the stated effect level (i.e., ERx). In other words, if $x > MDD%75$, then ERx is reasonable to use in risk assessment, since less than 25% of the studies fail to meet the threshold. If $x < MDD%75$, then ERx is not reasonable to use in risk assessment.

The results (Table 4) showed that in all cases (82/82), a 5% effect (ER05) cannot be reliably detected. In other words, the 75th percentile of the range of MDD% was at or below 5% in 0 out of 82 cases. It is expected to be possible to reliably detect the ER10, at a threshold of 75% of the time, in only 10/82...
cases (12%). Thus, neither of these measures of effect provides a sound basis for making risk-assessment conclusions from NTTP studies. The ER15 is expected to be reliably detected in 29 of 82 cases (35%). The ER25 is expected to be reliably detected in 67 of 82 cases (82%), indicating the suitability and robustness of this measure regardless of species, growth parameter, or study type.

Analysis of full dose–response data

The findings on the size of the effect that could be estimated based on the control analysis (discussed in the section Variability analysis for control data) were largely consistent with the results of fitting regression models to the datasets for which full dose–response data were available (see also Supplemental Information Attachment 1). The value of MDD% provides information on the effect size that can be distinguished from noise. When a regression model is fit to a data set, the MDD% provides an approximation to the effect size for which a reliable estimate can be obtained. It should be understood that once a regression model is fit to a data set, it is possible to estimate ERx for any value of x in the range 0 < x < 100. However, not all such estimates are useful. An estimated ERx is not useful if it requires extrapolation far above the largest tested application rate or far below the smallest application rate, nor is an estimate useful if its confidence interval is extremely wide. Further discussion of how to assess the quality of ECx estimates is given in Annex 6 of OECD TG 210 (OECD 2013). While that guideline is for fish, statistical annexes 5 and 6 are general. For the NTTP studies discussed here, with only rare exceptions, the common ratio of application rates to the next largest application rate is 2, 3, or 4. As should be expected (and was observed), there were many datasets for which no acceptable regression model was found, because in many instances, there was insufficient toxicity of the pesticide to result in a dose–response effect for each of the species tested with that substance.

An example of the dose–response analysis results is given for the shoot height growth response for the SE type of study. Of the 387 full dose–response datasets for SE shoot height, there were 365 (94%) for which ER05 could not be well estimated. Of those 365, there were 12 (3%) in which MDD% ≤ 5. Of the 22 datasets for which ER05 could be estimated, there were 4 (18%) in which MDD% exceeded 10.

Estimation of EC10 was more feasible. There were 295 of the 387 (76%) datasets for which EC10 could not be well estimated. Of these 295, the MDD% indicated a 10% effect could be detected in 74. Of the 92 studies for which an acceptable ER10 could be estimated, MDD% exceeded 20 in 10 studies (11%). Estimation of EC15 was slightly more feasible in that 284 of the 387 studies (73%) failed to produce an acceptable ER15. Of those 284, the MDD% indicated a 15% effect could be detected in 155 studies. Of the 103 studies where an ER15 could be well estimated, the MDD% exceeded 30 in 4 studies (4%).

Finally, ER25 could be well estimated in 279 of the 387 studies (72%). Of these 279, the MDD% was less than 25 in 241 studies (86%). Of the 108 studies where ER25 could not be well estimated, MDD% exceeded 30 in 5 studies (5%).

These results using the example of shoot height for SE indicate that estimation of EC05 is usually not possible and in cases when it is possible, it was unusual for MDD% to exceed 10. The success of estimating ERx was higher for higher effect sizes (i.e., values of x). When ERx could be well estimated, it was unusual for MDD% to exceed 2x. Thus, it is unlikely to be able to estimate ERx when MDD% exceeds x by more than a factor of 2. This is consistent with the theory (outlined in Supplemental Information Attachment 1) for nonlinear models.

Another aspect to be considered is how the data should be analyzed. The questions addressed in this regard are whether the raw data or a log-transform of the data should be analyzed. The normality and variance homogeneity of the data across studies for each plant species was assessed under each choice. If a log-transform is used, then it is preferable to calculate Log(1 + response), where response is shoot height or dry weight, to avoid negative values and distortions that can arise for responses less than 1, especially very small values that generate extremely large-magnitude negative logarithms if Log(response) were used instead. Both visual inspection of histograms of residuals from ANOVA and formal tests of normality (Shapiro–Wilk) and variance homogeneity (Levene) were used in this assessment.

### Table 4. Ability to reliably estimate various ERx values

| Test type                  | Growth response parameter | Allowing reliable estimate of ER05 | Allowing reliable estimate of ER10 | Allowing reliable estimate of ER15 | Allowing reliable estimate of ER25 |
|----------------------------|---------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|
| Seedling emergence         | dry weight                | 0/21                              | 0/21                              | 0/21                              | 9/21                              |
|                            | shoot height              | 0/21                              | 0/21                              | 8/21                              | 20/21                             |
| Vegetative vigor           | dry weight                | 0/20                              | 1/20                              | 4/20                              | 18/20                             |
|                            | shoot height              | 0/20                              | 9/20                              | 17/20                             | 20/20                             |
| Total                      |                           | 0/82                              | 10/82                             | 29/82                             | 67/82                             |
| Percent of total           |                           | 0%                                | 12%                               | 35%                               | 82%                               |

*Number (out of total) in which the 75th quantile of the MDD% (MDD%75 values) were less than the stated effect level (i.e., ERx).
Potential sources of variability in control data

Few clear trends were seen regarding the effect of specific study attributes on control variation. In part this was because there were too few observations for each combination of species, growth response, and study type with the attribute in question. Even when the data were simplified by listing controls as negative, adjuvant, or solvent without specifying the adjuvant or solvent, there were still no clear consistent differences. When fertilizer was restricted to being specified as “none,” “fertilizer,” and “nutrient solution” without specifying the type of fertilizer or nutrient, there was a general finding for SE studies of increased shoot height and dry weight when either was used compared to “none.” However, this was not observed in VV studies. Similarly, when soil type was restricted to “natural,” “blended,” or “artificial,” there was a tendency in SE studies for the measured responses to be greater in natural soil, but no such tendency was observed in VV studies. There were no clear differences across quarters (study time of year), but this analysis was complicated by numerous studies for which quarter was not specified.

CONCLUSIONS

Statistical differences of treatment groups from control groups are often a product of inherent variability, not phytotoxicity of the test substance. Inherent control variability is of particular concern because of the requirement to generate a NOER or, in its absence, an ER05 for each study type for use in endangered species assessments of pesticides. USEPA begins the process of conducting endangered species assessments for pesticides and, in the document *Interim Approaches for National-Level Pesticide Endangered Species Act Assessments Based on the Recommendations of the National Academy of Sciences April 2003 Report* (USEPA 2013), states how data from NTTP tests are used. The “concentration equal to the lowest value among the NOAEC and EC05 values from the available [SE] and [VV] studies” is used to establish the overlap of the action area with the range and critical habitats of endangered species to determine if a pesticide may affect threatened or endangered species and assess the potential for direct effects on terrestrial plants. This is an important facet of regulatory policy, because it means that if they provide the lowest value, flawed endpoints derived from overly variable tests will be perpetuated in future risk assessments.

As demonstrated by this data analysis, with the current NTTP test designs and implementation, it will often be impossible to reliably estimate an ER05, and it will rarely be possible to reliably estimate an ER10. In other words, estimated ER05 and ER10 values can be expected to represent responses that are not statistically significantly different from the controls and/or the confidence intervals for these endpoints will be quite large and generally include zero, such that the true value is highly uncertain. However, the ER25 can be reliably estimated in most cases. The MDD% was used as a suitable representation of the percent difference between the means of a treatment and the control that must exist to detect a statistically significant effect, and the median values of MDD% ranged from 3.5% to 39.5%, depending on the type of study, growth response parameter, and test species. The results of fitting regression models to the entire dose response (done for approximately 85% of the data) were consistent with the simpler control analysis.

Recent recommendations for risk assessments to consider ER05s in the place of nondefinitive NOERs (such as for assessments of imperiled species) should be approached with caution. Extrapolating to an ER05 without appropriate consideration of the statistical power of the test may result in inaccurate estimations of risk based on growth effect levels for which significant effects cannot be demonstrated.

Options for improving ERx estimation for small values of x in NTTP tests are limited. Increasing replication or increasing the number and spacing of test rates are difficult to implement, because most studies already use the maximum number of replicates and rates that can be accommodated in a greenhouse. The path forward may be to accept the ER25, which can be reliably estimated, as the appropriate test endpoint and explore alternative approaches when using the ER25 in risk assessments. Further research is needed in this area before making recommendations.

An important point to make about the analysis and conclusions presented herein is that the focus was solely on statistical significance. Biological relevance (i.e., the meaning of a certain percent effect on a growth response for risk assessment) is a separate and important consideration. Because the effects being observed in NTTP tests are sublethal effects on growth responses (e.g., shoot height and dry weight), consideration should be given to the potential for recovery in the absence of the test substance as well as what degree of effect to growth will actually cause a population-level effect.

Acknowledgment—We would like to thank the CropLife America members who provided data for use in this analysis, including DuPont, BASF, NovaSource/Tessenderlo Kerley, Inc., Syngenta, Bayer CropScience, Nufarm, Metalaxyl Task Force, FMC Corporation, Dow AgroSciences, the Industry Task Force II on 2,4-D Research Data, and ISK Biosciences Corporation. JW Green, D Edwards, K Henry, R Brain, B Glenn, N Ehresman, T Kung, K Ralston-Hooper, F Kee, and S McMaster are employed by CropLife America members. We would also like to thank the CropLife America Environmental Risk Assessment Committee and Ecotoxicology Working Group for supporting this project. Funding for aspects of this project was provided by CropLife America.

Disclaimer—The peer review of this article was managed by the Editorial Board without the involvement of R Brain.

Data Accessibility—The data for the control replicates for each study are provided in the Supplemental Information, Attachment 2.

SUPPLEMENTAL DATA

Supplemental Information Attachment 1. Derived relationship between MDD% and MSEff; details regarding regression models used and reliability criteria.

Supplemental Information Attachment 2. NTTP data.
REFERENCES

Brain RA, Bestari KJ, Sanderson H, Hanson ML, Wilson CJ, Johnson DJ, Sibley PK, Solomon KR. 2005. Aquatic microcosm assessment of the effects of tylosin on Lemna gibba and Myriophyllum spicatum. Environ Pollut 133(3):389–401.

Brain RA, Johnson DJ, Richards SM, Hanson ML, Sanderson H, Lam MW, Young C, Mabury SA, Sibley PK, Solomon KR. 2004. Microcosm evaluation of the effects of an eight pharmaceutical mixture to the aquatic macrophytes Lemna gibba and Myriophyllum spicatum. Aquat Toxicol 70(1):23–40.

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.

Davy M, Petrie R, Smrckov J, Kuchnicki T, Francois D. 2001. Proposal to update non-target plant toxicity testing under NAFTA: Scientific Advisory Panel briefing. US Environmental Protection Agency and Health Canada. 164 p.

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

[EFSA] European Food Safety Authority. 2014. Scientific opinion addressing the state of the science on risk assessment of plant protection products for non-target terrestrial plants, EFSA Panel on Plant Protection Products and their Residues (PPR). EFSA Journal 12(7):3800.

Fox DR. 2011. Response to Landis and Chapman (2011). Integr Environ Assess Manag 8:4.

Green JW. 2016. Statistical challenges in ecotoxicology. Environ Toxicol Chem 34:2437.

Green JW. 2016. Issues with using only regression models for ecotoxicology studies. Integr Environ Assess Manag 12(1):198–199.

Green JW, Springer TA, Staveley JP. 2012. The drive to ban the NOEC/LOEC in favor of ECx is misguided and misinformed. Integr Environ Assess Manag 8(2):242–249.

Hanson ML, Sanderson H, Solomon KR. 2003. Variation, replication, and power analysis of Myriophyllum spp. microcosm toxicity data. Environ Toxicol Chem 22(6):1318–29.

Harcum JB, Dressing SA. 2015. Technical memorandum #3: Minimum detectable change and power analysis, October 2015. Developed for U.S. Environmental Protection Agency. Fairfax (VA): Tetra Tech, Inc. 10 p. [cited 2018 March 2]. https://www.epa.gov/sites/production/files/2015-10/documents/tech_memo_3_oct15.pdf

Hogg RV, Tanis EA. 1986. Probability and statistical inference. 3rd ed. New York (NY): Macmillan. 424 p.

Holst RW, Ellwanger TC. 1982. Pesticide assessment guidelines subdivision J hazard evaluation: non-target plants. Washington (DC): US Environmental Protection Agency. EPA-540/9-82-020. p 1–55.

Jager T. 2012. Bad habits die hard: The NOEC’s persistence relies poorly on ecotoxicology. Environ Toxicol Chem 31:228–229.

Knauer K, Vervliet-Scheebaum M, Dark RJ, Maund SJ. 2006. Methods for assessing the toxicity of herbicides to submersed aquatic plants. Pest Manag Sci 62(8):715–22.

Kraufvelin P. 1998. Model ecosystem replicability challenged by the “soft” reality of a hard bottom mesocosm. J Exp Mar Biol Ecol 222(1):247–67.

Landis WG, Chapman PM. 2011. Well past time to stop using NOELs and LOELs. Integr Environ Assess Manag 7:vii–viii.

Meinert CL. 2012. Clinical trials: design, conduct and analysis. New York (NY): Oxford University Press. 720 p.

[OECD] Organisation for Economic Co-operation and Development. 2014. Current approaches in the statistical analysis of ecotoxicity data: A guidance to application. [Annexes to this publication exist as a separate document]. OECD Series on Testing and Assessment, No 54. Paris (FR): OECD Publishing. 147 p. [cited 2018 April 24]. https://doi.org/10.1787/9789264085275-en

[OECD] Organisation for Economic Co-operation and Development. 2013. Fish, early-life stage toxicity test. OECD guidelines for the testing of chemicals. Sect 2, No 210. Paris (FR): OECD Publishing. 24 p. [cited 2018 April 24]. https://doi.org/10.1787/9789264023785-en

[OECD] Organisation for Economic Co-operation and Development. 2006a. Terrestrial plant test: Seedling emergence and seedling growth test. OECD guidelines for the testing of chemicals. Sect 2, No 208. Paris (FR): OECD Publishing. 21 p. [cited 2018 April 24]. https://doi.org/10.1787/9789264070066-en

[OECD] Organisation for Economic Co-operation and Development. 2006b. Terrestrial plant test: Vegetative vigour test. OECD Guidelines for the Testing of Chemicals, Sect 2, No 227. Paris (FR): OECD Publishing. [cited 2018 April 24]. https://doi.org/10.1787/9789264057295-en

Sanderson H, Laird B, Brain R, Wilson CJ, Solomon KR. 2009. Detectability of fifteen aquatic micro/mesocosms. Ecotoxicology 18(7):838–45.

Stephan CE. 1977. Methods for calculating an LC50. In: Mayer FL, Hamelink JL, editors. Aquatic toxicology and hazard evaluation. ASTM STP 634. Philadelphia (PA): American Society for Testing and Materials. p 65–84.

Trevan JW. 1927. The error of determination of toxicity. Proc R Soc Lond B Biol Sci 101:483–514.

[USEPA] US Environmental Protection Agency. 2012a. Ecological effects test guidelines. OCSP 850.4100: Seedling emergence and seedling growth. US EPA, Office of Chemical Safety and Pollution Prevention (7101). 32 p. EPA 712-C-012.

[USEPA] US Environmental Protection Agency. 2012b. Ecological effects test guidelines. OCSP 850.4150: Vegetative vigor. US EPA, Office of Chemical Safety and Pollution Prevention (7101). 20 p. EPA 712-C-012.

[USEPA] US Environmental Protection Agency. 2012c. Ecological effects test guidelines. OCSP 850.4100: Seedling emergence and seedling growth. US EPA, Office of Chemical Safety and Pollution Prevention (7101). 32 p. EPA 712-C-012.

[USEPA] US Environmental Protection Agency. 2012d. Ecological effects test guidelines. OCSP 850.4000: Background and special considerations—tests with terrestrial wildlife. US EPA, Office of Chemical Safety and Pollution Prevention (7101). 29 p. EPA 712-C-026.

[USEPA] US Environmental Protection Agency. 2013. Interim approaches for national-level pesticide Endangered Species Act assessments based on the recommendations of the national academy of sciences April 2003 report. US EPA. 10 p. https://www.epa.gov/sites/production/files/2015-07/documents/interagency.pdf

van Dam RA, Harford AJ, Warner 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(2):242–245.

Wieczorek MV, Bakanov N, Lagadic L, Bruns E, Schulz R. 2017. Response and recovery of the macrophytes Elodea canadensis and Myriophyllum spicatum following a pulse exposure to the herbicide ifosulfuron-sodium in outdoor stream mesocosms. Environ Toxicol Chem 36(4): 1090–1100.