Supplementary materials for this article are available at http://pubs.amstat.org/toc/jabes/14/4.

Testing for the Equality of EC50 Values in the Presence of Unequal Slopes With Application to Toxicity of Selenium Types

Daniel R. JESKE, Huaying Karen XU, Todd BLESSINGER, Peter JENSEN, and John TRUMBLE

The likelihood ratio test (LRT) for the equality of EC50 values using a probit model that has parallel slopes is implemented in a variety of software packages. A preliminary LRT can be used to ascertain the plausibility of parallel slopes. Testing for equal EC50 values is not as straightforward if the preliminary test rejects that the slopes are equal or, equivalently, if a practitioner would rather not deal with the implications on the size of the test in the presence of the preliminary test. An LRT for testing equal EC50 values is not available in software packages for the case of arbitrary slopes. In this article, we describe a simple and effective algorithm for implementing the LRT procedure in this case. We also derive a quadratic form test procedure for the same hypothesis and compare the two tests (size and power) in the context of our application that deals with comparing the toxicity of four different types of selenium. The R-code is available as supplemental material online.

Key Words: Equality of EC50 values; Natural response parameter; Probit regression.

1. INTRODUCTION

The scientific literature in entomology, environmental science, and toxicology frequently use log-dose probit analyses of chemicals to determine effective concentrations (EC) of toxins for specific organisms. EC levels are the concentration levels needed to eliminate a specified level of the population. Of frequent interest is the EC50 level, which is the concentration level that kills 50% of the population. Comparative data on EC levels

Daniel R. Jeske is Professor of Statistics and Director, Statistical Consulting Collaboratory, Department of Statistics, University of California, Riverside, CA 92521 (E-mail: *daniel.jeske@ucr.edu*). Huaying Karen Xu is Associate Director, Statistical Consulting Collaboratory, Department of Statistics, University of California, Riverside, CA 92521 (E-mail: *karen.xu@ucr.edu*). Todd Blessinger is Research Scientist, Federal Drug Administration, Rockville, MD 20855 (E-mail: *todd.blessinger@fda.hhs.gov*). Peter Jensen is Post-doc, Department of Entomology, University of Maryland, College Park, MD 20742 (E-mail: *pjensen@umd.edu*). John Trumble is Professor of Entomology, University of California, Riverside, CA 92521 (E-mail: *john.trumble@ucr.edu*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 14, Number 4, Pages 469–483
DOI: 10.1198/jabes.2009.07088

historically have been used for ranking toxicants, particularly pesticides (e.g., see Brown 1958 for an early reference). Rankings are currently used to determine the signal words (e.g., danger, warning, etc.) required by the U.S. Environmental Protection Agency (EPA) on all pesticide labels. Information on effective concentrations can also be used to compare specific forms (or formulations) of individual toxicants as well as combinations of toxicants. This information can provide insight into the potential additive, antagonistic, or synergistic effects of chemicals and combinations of chemicals (e.g., Van Gestel and Hensbergen 1997). Finally, in risk assessment studies, such comparative toxicological data can be used to estimate noeffect levels for further testing, or to help decision making efforts when production or release of two or more toxicants is contemplated.

Ranking toxins based on EC values is more complicated when the dose-to-response curves for the toxins have unequal slopes. In this case, the ranking results depend on which EC level is used. This complication has been recognized in the toxicology literature, but it is still common practice to use the EC50 level for ranking purposes (see Van Birgelen et al. 1995 and Zeiger et al. 2001). One reason for continuing to use EC50 values for ranking in this case is that they are the EC levels that can be estimated with the highest precision. Lacking in the literature to date, but discussed in this article, is a suitable precursor test (to the ranking) of whether the underlying EC50 values are all equal when the toxins have different slopes.

The data from a typical toxicology experiment take the form of independent observations $\{Y_{ij} : 1 \le i \le I, 1 \le j \le J_i\}$, where Y_{ij} denotes the number of organisms that were killed when exposed to the jth concentration level of the ith toxin. Here, the notation indicates that I toxins are under study and the ith toxin utilized J_i different concentration levels. The probability model for Y_{ij} is binomial with sample size n_{ij} and probability of death p_{ij}. Here, n_{ij} represents the number of organisms exposed to the jth concentration level of the ith toxin.

A traditional toxicology model for the Y_{ij} observations extends the binomial model by relating the p_{ij} values to the concentration levels of the toxins through a so-called probit model of the form $\Phi^{-1}(p_{ij}) = \alpha_i + \beta_i x_{ij}$, where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function of the standard normal distribution, x_{ij} is the (known) jth concentration level that was used for the ith toxin, and (α_i, β_i) are unknown toxin-dependent intercept and slope parameters. An equivalent way to express the probit model is $p_{ij} = \Phi(\alpha_i + \beta_i x_{ij})$. A generalized probit model that incorporates death by natural causes is $p_{ij} = \delta + (1 - \delta)\Phi(\alpha_i + \beta_i x_{ij})$, where δ represents the probability an organism died in the absence of any toxin (i.e., death by an otherwise natural cause). Often δ is estimated by exclusively using observations from a control experiment where the concentration level is zero, and then interpreting the estimated δ as a known value. Throughout the remainder of this article, we adopt this practice. Under the generalized probit model, EC levels are defined with respect to the subpopulation of organisms that escape death by natural causes. Hence, the EC level of the ith toxin that kills $x\%$ of the population is $(\Phi^{-1}(x/100) - \alpha_i)/\beta_i$. In particular, we let the EC50 values be represented by $R_i = -\alpha_i/\beta_i$.

Our toxicology experiments were performed to compare the toxicity on flies of four different forms of selenium. Selenium is a metalloid with several valent states, each of

Table 1. Data from toxicology experiments with selenium.

Selenium type	Conc. x_{ij}	Samples, n_{ij}	Deaths Y_{ij}	Selenium type	Conc. x_{ij}	Samples, n_{ij}	Deaths Y_{ij}
1	0	151	3	2	0	141	2
1	100	146	40	2	100	153	30
1	200	116	31	2	200	142	59
1	300	159	85	2	300	139	82
1	400	150	102	2	400	154	62
1	500	140	112	2	500	155	85
3	0	137	4	4	0	152	3
3	5	106	0	4	5	152	7
3	25	63	11	4	25	150	11
3	50	145	22	4	50	153	45
3	100	127	31	4	100	125	74
3	200	140	105				
3	400	172	166				
3	800	188	188				

which are found in the environment at significant concentrations. Historically, toxicological data have been collected only for *total* selenium levels, as it was not possible to detect the different forms of selenium in organisms or the environment. Detection is now possible with the emergence of new techniques and technology and thus, the important problem of characterizing the toxicological profiles of the most prominent forms of selenium can be addressed. The data in Table 1 resulted from our experiments with four different forms of selenium (selenate, selenite, selenomethionine, and selenocysteine). We subsequently identify the four forms of selenium as types 1–4.

The four observations from Table 1 at zero concentration can be pooled to obtain $\delta = 0.021$ as the probability of death by natural causes. Maximum likelihood estimates (MLEs) and their standard errors (shown in parentheses) of the slope and intercept of the generalized probit models for each selenium type are shown in Table 2, along with the MLEs and asymptotic confidence intervals (derived using the delta method) of the corresponding EC50 values. The plausibility of the probit model is illustrated by the linearity in

Table 2. MLEs, standard errors, and EC50 confidence intervals for selenium data.

			EC50 value	
Selenium type	MLE intercept	MLE slope	MLE	90% conf. interval
1	−5.26	0.948	5.55	(5.45, 5.66)
	(0.54)	(0.095)	(0.064)	
2	−3.23	0.540	5.99	(5.32, 6.64)
	(0.50)	(0.088)	(0.40)	
3	−6.64	1.37	4.84	(4.63, 5.04)
	(0.52)	(0.102)	(0.13)	
4	−5.60	1.27	4.42	(4.10, 4.73)
	(0.63)	(0.153)	(0.19)	

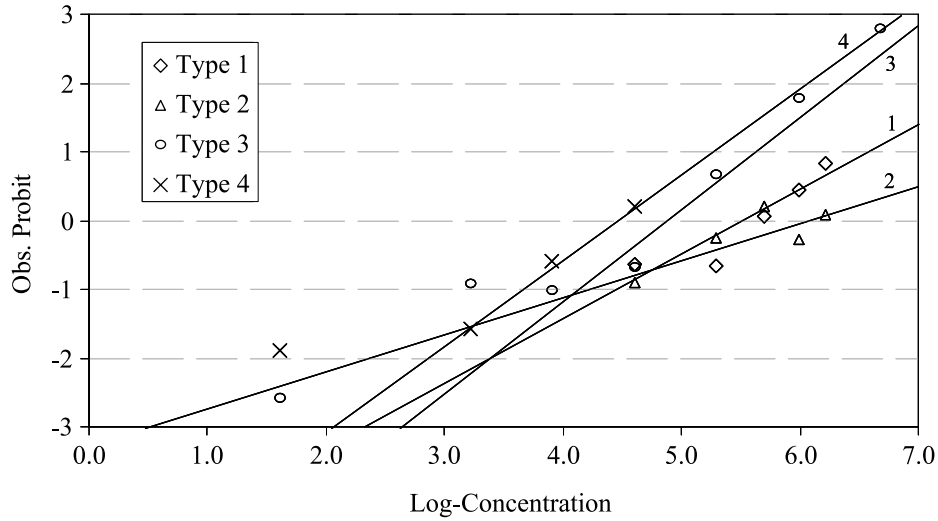

Figure 1. Observed probit plots for selenium experiments.

Figure 1 which is a plot of the empirical probits, $\Phi^{-1}\{[(Y_{ij}+0.5)/(n_{ij}+1)-\delta]/(1-\delta)\}$, versus $\ln(x_{ij})$ for the nonzero concentrations and for each selenium type. (Note that we have defined the empirical probits using a modified estimator of p_{ij} that accommodates the two extreme cases where Y_{ij} is zero and n_{ij}.) Only the two points at the lowest concentration level show appreciable deviations from the fitted (MLE) lines (which are labeled 1–4 to map them to the selenium types). These points correspond to concentration level $x = 5$ that was used for selenium types 3 and 4, a concentration that is arguably too low to have an effect on mortality. In terms of fitting the model, these observations are not influential on either the parameter estimates or any of the inferences subsequently drawn.

While it is common to find MLEs of EC50 values from probit analyses in the toxicology literature, it is a little less common to find a hypothesis test procedure to determine whether estimated differences between EC50 values are statistically significant. In cases where inferences of this nature have been drawn, the use of nonoverlapping confidence limits for the EC50 values has frequently been used as an indication of significant differences (e.g., Abot et al. 1995; Liu et al. 2003). Fieller's theorem (see Cox 1990) is typically used to obtain the individual EC50 confidence limits. Use of nonoverlapping confidence limits for EC50 values to declare significant differences results in a conservative test. Schenker and Gentleman (2001) and Payton, Greenstone, and Schenker (2003) are good references for further discussion on the widespread use of the overlapping confidence intervals technique and its associated limitations.

An LRT for testing the equality of EC50 values is well known (e.g., Sokal and Rholf 1969) and is implemented in many statistical analysis software packages. However, to date implementations of the LRT assume equality of the slopes in the probit model. If the slopes are not equal, an LRT test is not directly available in any software packages that we are aware of; we know of no applications where it has otherwise been used to test the equality of EC50 values. As a result, we were motivated to develop a simple and effective algo-

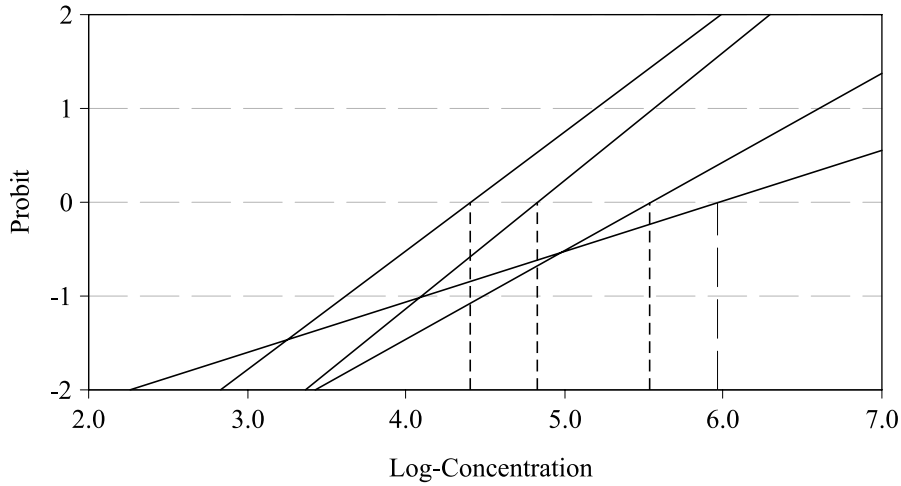

Figure 2a. Probit regressions with different slopes and different EC50 values (x-intercepts).

rithm for computing the LRT. Our purpose in this article is to describe it in the belief it will be beneficial to other practitioners interested in testing for equal EC50 values, as a precursor analysis to ranking, without making the assumption of equal slopes. To gain insight into the performance of the LRT procedure (i.e., size and power), we compare it with a quadratic form test procedure which has the appearance of being more straightforward from a computational point of view.

More formally, we consider the generalized probit model, with δ assumed known, and develop the LRT and quadratic form tests for $H : \alpha_1/\beta_1 = \cdots = \alpha_I/\beta_I$ vs. $K : \sim H$. Figures 2a and 2b provide a visual contrast between H and K, showing that under K the different probit regression lines potentially all have different x-intercepts while under H all of the lines have the same x-intercept which is the common EC50 value. A reduced

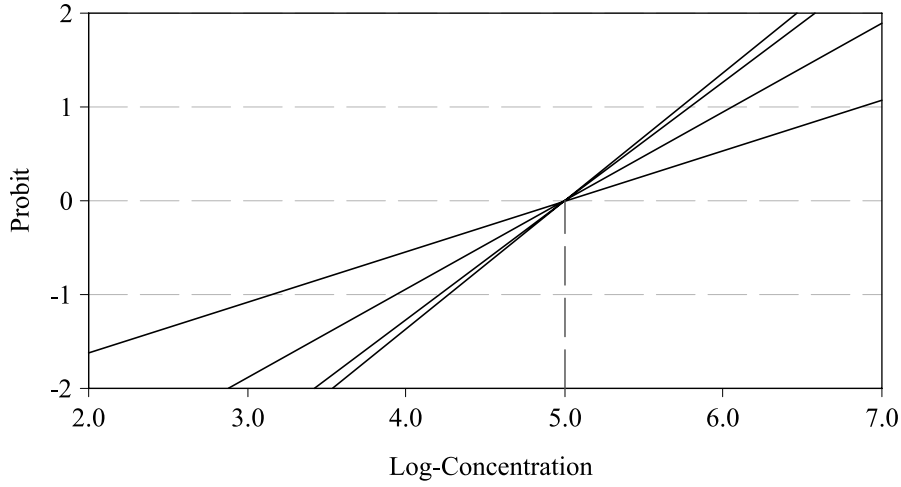

Figure 2b. Probit regressions with different slopes and a common EC50 value (the x-intercept).

model, known as the parallel-slope model, adds a constraint to the probit model of the form $C: \beta_1 = \cdots = \beta_I$. Under this constraint, H reduces to the hypothesis that the x-intercepts of the probit regression lines are all equal. As mentioned above, the LRT of H under the parallel-slope model is a well-known test and is implemented in commercial statistical software packages. In the same packages, it is also possible to obtain an LRT for testing if the constraint C is valid, and this could be done as a preliminary test to the LRT of H vs. K (e.g., Van Gestel and Hensbergen 1997). The problem we confront in this article is how to test H vs. K when the constraint C is not valid, either because the preliminary test of its validity was rejected or because a practitioner is otherwise unwilling to adopt the parallel-slope model.

The rest of this article is organized as follows. In Section 2, we develop the LRT and the quadratic form test procedures of $H: \alpha_1/\beta_1 = \cdots = \alpha_I/\beta_I$ versus $K: \sim H$. In Section 3, we use a simulation study to investigate the adequacy of the asymptotic null distributions of the two test procedures and demonstrate the need to calibrate the size of the quadratic form test. The two test procedures are illustrated in Section 4 with the experimental data collected from our own selenium toxicity experiments. The power of the two test procedures is analyzed in Section 5 with a simulation study, where we find that the performance of the calibrated quadratic form test procedure is equivalent to the LRT test procedure. We conclude with a summary in Section 6 that includes some comments about how the test procedure can be extended to related contexts.

2. TEST FOR EQUALITY OF EC LEVELS

2.1 LIKELIHOOD RATIO TEST

With $\boldsymbol{\theta} = (\alpha_1, \beta_1, \ldots, \alpha_I, \beta_I)'$, the null hypothesis can be expressed as $H: \mathbf{g}(\boldsymbol{\theta}) = 0$, where $\mathbf{g}(\boldsymbol{\theta})$ is a $(I-1) \times 1$ vector whose elements are $g_i(\boldsymbol{\theta}) = \alpha_i/\beta_i - \alpha_I/\beta_I$ $(i = 1, \ldots, I-1)$. The likelihood function for $\boldsymbol{\theta}$ based on the observations $\{Y_{ij} : 1 \le i \le I, 1 \le j \le J_i\}$ can be written as (recall that δ is being regarded as a known constant):

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{I} \prod_{j=1}^{J_i} [\delta + (1-\delta)\Phi(\alpha_i + \beta_i x_{ij})]^{y_{ij}} [1 - \Phi(\alpha_i + \beta_i x_{ij})]^{n_{ij} - y_{ij}}.$$

The LRT statistic is $\Lambda = -2\log[L(\tilde{\boldsymbol{\theta}})/L(\hat{\boldsymbol{\theta}})]$, where $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ denote the restricted maximum likelihood estimate (under H) and the unrestricted maximum likelihood of $\boldsymbol{\theta}$, respectively. Under the null hypothesis H, the asymptotic chi-square distribution of Λ is chi-square with $I-1$ degrees of freedom.

Calculation of the denominator of Λ, via $\hat{\boldsymbol{\theta}}$, is straightforward since the standard software packages can be used on the data individually for each toxin to find the MLE of the intercept and slope parameters, say $\{(\hat{\alpha}_i, \hat{\beta}_i)\}_{i=1}^{I}$. Calculation of the numerator of Λ, via $\tilde{\boldsymbol{\theta}}$, is a little more challenging as is any restricted optimization relative to an unrestricted optimization. However, the iterative Lagrangian algorithm outlined in Henk (1985) provided a relatively simple and very effective solution. In Henk's algorithm, the iterates for $\tilde{\boldsymbol{\theta}}$ were given by:

$$\tilde{\boldsymbol{\theta}}_{k+1} = \tilde{\boldsymbol{\theta}}_k + \mathbf{P}(\tilde{\boldsymbol{\theta}}_k)\mathbf{l}(\tilde{\boldsymbol{\theta}}_k) - [\mathbf{I} - \mathbf{P}(\tilde{\boldsymbol{\theta}}_k)\mathbf{B}(\tilde{\boldsymbol{\theta}}_k)]\mathbf{G}^{+}(\tilde{\boldsymbol{\theta}}_k)\mathbf{g}(\tilde{\boldsymbol{\theta}}_k), \qquad k = 0, 1, \ldots,$$

where $\mathbf{l}(\boldsymbol{\theta})$ is the $2I \times 1$ vector of derivatives of the log-likelihood with respect to the components of $\boldsymbol{\theta}$, $\mathbf{G}(\boldsymbol{\theta})$ is the $(I-1) \times 2I$ matrix whose (i, j)th element is $\partial g_i(\boldsymbol{\theta})/\partial\theta_j$, $\mathbf{G}^+(\boldsymbol{\theta})$ is the Moore–Penrose generalized inverse of $\mathbf{G}(\boldsymbol{\theta})$, $\mathbf{B}(\boldsymbol{\theta})$ is the $2I \times 2I$ matrix whose (i, j)th element is the expected value (under the full model) of $-\partial^2\mathbf{l}(\boldsymbol{\theta})/\partial\theta_j\,\partial\theta_i$, $\mathbf{I}$ is $2I \times 2I$ identity matrix, and $\mathbf{P}(\boldsymbol{\theta})$ is the Moore–Penrose generalized inverse of $\mathbf{I} - \mathbf{G}^+(\boldsymbol{\theta})\mathbf{G}(\boldsymbol{\theta})$.

As with any iterative optimization algorithm, a critical question is what to choose for the initial starting value $\hat{\boldsymbol{\theta}}_0$. In our particular context, we see that H implies $\alpha_i = -R\beta_i$ $(i = 1, \ldots, I)$, where R is the common EC50 value. Fitting a nointercept least-squares lines through the points $\{(-\hat{\beta}_i, \hat{\alpha}_i)\}_{i=1}^I$ produces a value $\tilde{R}$ that is a crude, but useful, first guess at the restricted maximum likelihood estimate of the common value of the common EC50 value. Consequently, we define $\tilde{\boldsymbol{\theta}}_0 = (-\tilde{R}\hat{\beta}_1, \hat{\beta}_1, -\tilde{R}\hat{\beta}_2, \hat{\beta}_2, \ldots, -\tilde{R}\hat{\beta}_I, \hat{\beta}_I)'$. Our experience shows that, in addition to being simple, this choice for the starting value is quite effective.

2.2 QUADRATIC FORM TEST

We begin by noting that for each $i = 2, \ldots, I$, the ith EC50 value will be equal to the first EC50 value if and only if $\alpha_1\beta_i - \alpha_i\beta_1 = 0$. Setting $\phi_i = \alpha_1\beta_i - \alpha_i\beta_1$ and $\boldsymbol{\phi} = (\phi_2, \ldots, \phi_I)'$, it is easily seen that testing H vs. K is equivalent to testing $H^*: \boldsymbol{\phi} = 0$ vs. $K^*: \sim H^*$. Let $\hat{\boldsymbol{\theta}}_i = (\hat{\alpha}_i, \hat{\beta}_i)'$ and $\hat{\phi}_i = \hat{\alpha}_1\hat{\beta}_i - \hat{\alpha}_i\hat{\beta}_1$ denote the MLEs of $\boldsymbol{\theta}_i = (\alpha_i, \beta_i)'$ and ϕ_i, respectively.

Expanding each $\hat{\phi}_i$ in a Taylor series about the point $(\boldsymbol{\theta}_1', \boldsymbol{\theta}_i')'$ leads to $\hat{\boldsymbol{\phi}} - \boldsymbol{\phi} \approx \mathbf{A}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, where $\mathbf{A}(\boldsymbol{\theta})$ is a $(I-1) \times 2I$ matrix whose ith row contains the first-order partial derivatives of ϕ_i, with respect to the elements of $\boldsymbol{\theta}$. Denoting the inverse of the expected Fisher information matrix of $\hat{\boldsymbol{\theta}}_i$ by $\boldsymbol{\Sigma}_i(\boldsymbol{\theta}_i)$, it follows that an approximation to the variance–covariance matrix of $\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}$ is then:

$$\begin{aligned}\boldsymbol{\Sigma}(\boldsymbol{\theta}) &= \mathbf{A}(\boldsymbol{\theta})\,\mathrm{var}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\mathbf{A}'(\boldsymbol{\theta}) \\ &= \mathbf{A}(\boldsymbol{\theta})\,\mathrm{diag}[\boldsymbol{\Sigma}_i(\boldsymbol{\theta}_i)]_{i=1}^I\mathbf{A}'(\boldsymbol{\theta}) \\ &= \mathrm{diag}[\boldsymbol{\theta}_1'\mathbf{U}\boldsymbol{\Sigma}_i(\boldsymbol{\theta}_i)\mathbf{U}'\boldsymbol{\theta}_1]_{i=2}^I + [\boldsymbol{\theta}_i'\mathbf{U}\boldsymbol{\Sigma}_1(\boldsymbol{\theta}_1)\mathbf{U}'\boldsymbol{\theta}_j]_{i,j=2}^I, \end{aligned} \tag{2.1}$$

where $\mathbf{U} = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$. Let $S(\boldsymbol{\theta})$ denote Equation (2.1), but with each $\boldsymbol{\Sigma}_i(\boldsymbol{\theta}_i)$ replaced by the inverse of the corresponding observed Fisher information matrix, say $S_i(\boldsymbol{\theta}_i)$. Then a large sample theory approach to testing H^* vs. K^* would be to use the statistic $T_1 = \hat{\boldsymbol{\phi}}' S^{-1}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\phi}}$, which under H^* has an asymptotic chi-square distribution with $I-1$ degrees of freedom. A nominal size γ test based on this large sample test statistic rejects H^* if and only if $T_1 > \chi^2_{I-1,\gamma}$. We note that T_1 is a particular implementation of a Wald test statistic for the case of a nonlinear hypothesis.

The expansion leading to Equation (2.1) results in approximating the variance of $\hat{\phi}_i$ by a quantity that is too small. This can be seen by using the law of total variance and the asymptotic normal distribution of $\hat{\boldsymbol{\theta}}$ to obtain:

$$\mathrm{var}(\hat{\phi}_i) = E\{\mathrm{var}(\hat{\phi}_i|\hat{\boldsymbol{\theta}}_1)\} + \mathrm{var}\{E(\hat{\phi}_i|\hat{\boldsymbol{\theta}}_1)\}$$

$$
\begin{aligned}
&= E\{\hat{\alpha}_1^2 \operatorname{var}(\hat{\beta}_i) + \hat{\beta}_1^2 \operatorname{var}(\hat{\alpha}_i) - 2\hat{\alpha}_1\hat{\beta}_1 \operatorname{cov}(\hat{\alpha}_i, \hat{\beta}_i)\} \\
&\quad + \operatorname{var}\{\hat{\alpha}_1 E(\hat{\beta}_i) - \hat{\beta}_1 E(\hat{\alpha}_i)\} \\
&\approx \boldsymbol{\theta}_1' \mathbf{U}\boldsymbol{\Sigma}_i(\boldsymbol{\theta}_i)\mathbf{U}'\boldsymbol{\theta}_1 + \operatorname{tr}(\boldsymbol{\Sigma}_1(\boldsymbol{\theta}_1)\mathbf{U}\boldsymbol{\Sigma}_i(\boldsymbol{\theta}_i)\mathbf{U}') + \boldsymbol{\theta}_i' \mathbf{U}\boldsymbol{\Sigma}_1(\boldsymbol{\theta}_1)\mathbf{U}'\boldsymbol{\theta}_i, \qquad (2.2)
\end{aligned}
$$

where $\operatorname{tr}(\cdot)$ denotes the trace operator on matrices. Note that the (positive) trace term in Equation (2.2) is the correction relative to the underestimate provided by Equation (2.1). The approximations of the covariance terms $\operatorname{cov}(\hat{\phi}_i, \hat{\phi}_j)$ implied by Equation (2.1) agree with what would similarly be obtained by using the law of total covariance. The corrected variance terms in Equation (2.2) can naturally be incorporated to give a corrected approximation to the variance–covariance matrix of $\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}$ of the form:

$$
\begin{aligned}
\boldsymbol{\Sigma}_c(\boldsymbol{\theta}) &= \operatorname{diag}\left[\boldsymbol{\theta}_1' \mathbf{U}\boldsymbol{\Sigma}_i(\boldsymbol{\theta}_i)\mathbf{U}'\boldsymbol{\theta}_1 + \operatorname{tr}\{\boldsymbol{\Sigma}_1(\boldsymbol{\theta}_1)\mathbf{U}\boldsymbol{\Sigma}_i(\boldsymbol{\theta}_i)\mathbf{U}'\}\right]_{i=2}^{I} \\
&\quad + [\boldsymbol{\theta}_i' \mathbf{U}\boldsymbol{\Sigma}_1(\boldsymbol{\theta}_1)\mathbf{U}'\boldsymbol{\theta}_j]_{i,j=2}^{I}, \qquad (2.3)
\end{aligned}
$$

which in turn suggests $T_2 = \hat{\boldsymbol{\phi}}' S_c^{-1}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\phi}}$ as a modified test statistic, where $S_c(\boldsymbol{\theta})$ denotes Equation (2.3) with each $\boldsymbol{\Sigma}_i(\boldsymbol{\theta}_i)$ replaced by $S_i(\boldsymbol{\theta}_i)$. An alternative nominal size γ test of H^* vs. K^* rejects H^* if and only if $T_2 > \chi^2_{I-1,\gamma}$.

3. SIZE AND CALIBRATION

All three statistics Λ, T_1, and T_2 have asymptotic chi-square distributions with $I-1$ degrees of freedom. We utilize a small simulation study to investigate the adequacy of using the chi-square distribution as the appropriate null distribution. Design parameters for the study are motivated by the design parameters associated with our selenium experiments. Consequently, four toxicological treatments are considered and for each of them, concentration levels of 10, 25, 50, 100, 200, and 400 are used. In terms of the notation introduced in Section 1, we have $I = 4$, $J_1 = \cdots = J_4 = 6$. We take $n_{ij} = 30$ for all (i, j) as a suitably small value to test the adequacy of the asymptotic distribution. The underlying model is the generalized probit model described in Section 1, using log-concentration as x_{ij}.

For the size simulations, we used 5 as a common EC50 value (corresponding to roughly the center of the four observed selenium EC50 values), $\delta = 0.021$ (the observed value in the selenium experiments) and defined four equal sized intervals $B_1 = [0.25, 0.5625)$, $B_2 = [0.5625, 0.875)$, $B_3 = [0.875, 1.1875)$, and $B_4 = [1.1875, 1.5]$ which surround the range of slopes observed in our selenium application. For each interval B_j, we randomly sampled 20 slope vectors $(\beta_1, \beta_2, \beta_3, \beta_4)'$ and then for each slope vector 1000 simulated datasets were generated. For each generated dataset, the three test statistics Λ, T_1, and T_2 were computed and compared to the $\chi^2_{3,0.9}$ critical value to carry out nominal 10% Type I error tests. The average power for the 20 repetitions corresponding to different slope vectors is reported in Table 3 as a summary of the size associated with the slope interval B_j (and the common value of 5 for the EC50 values).

Table 3 suggests the asymptotic chi-square distribution is adequate for Λ for sample sizes as small as 30, but that small sample calibration is needed for T_1 and T_2. [We note that additional simulations (not reported here) confirm that by increasing the value of n_{ij},

Table 3. Simulated size (see Section 3) of nominal 10% tests of H_0 : equal EC50 values.

	Test statistic				
Interval for slopes	T_1	T_1 (calibrated)	T_2	T_2 (calibrated)	Λ
B_1	0.048	0.11	0.041	0.11	0.10
B_2	0.062	0.10	0.056	0.10	0.10
B_3	0.056	0.11	0.050	0.11	0.11
B_4	0.050	0.11	0.044	0.11	0.11

the size of the uncalibrated tests based on for T_1 and T_2 do, in fact, converge to the correct nominal size.] Appendix A illustrates how bootstrap calibrated p-values (e.g., Chernick 1999) can be obtained for T_1 or T_2 (generically denoted by T in the algorithm). The basic idea is to create a histogram of bootstrap values of the statistic using simulated datasets that are generated under a null model that is estimated from the observed dataset. In typical applications of bootstrap null distributions, the null model used to generate the simulated datasets is the model fit using the constrained (by H) MLE. An alternative, simpler approach, defined in step 3 of Appendix A, is to generate null datasets using the value $\tilde{\boldsymbol{\theta}}_0 = (-\tilde{R}\hat{\beta}_1, \hat{\beta}_1, -\tilde{R}\hat{\beta}_2, \hat{\beta}_2, \ldots, -\tilde{R}\hat{\beta}_I, \hat{\beta}_I)'$, the starting value associated with the Henk algorithm discussed in Section 2.1. Table 3 shows that bootstrap calibration of the p-values obtained from the tests based on T_1 or T_2 effectively corrects the size.

4. APPLICATION OF TESTS TO SELENIUM DATA

4.1 LIKELIHOOD RATIO TEST

Table 2 shows the unrestricted MLEs $\hat{\boldsymbol{\theta}}_i$ $(i = 1, \ldots, 4)$ that were obtained using PROC PROBIT in the SAS language. The standard chi-square LRT (with 3 degrees of freedom) for equal slopes is rejected based on a test statistic value of 48.46. Table 2 also shows the corresponding MLEs of the EC50 values along with 90% confidence intervals for the EC50 values. The EC50 confidence intervals were computed using the delta method. The value of $L(\hat{\boldsymbol{\theta}})$ is easily calculated.

To compute the numerator of the LRT statistic, Henk's algorithm was employed. Defining $u_{ij} = \alpha_i + \beta_i x_{ij}$, the derivatives needed for the evaluation of $\mathbf{I}(\boldsymbol{\theta})$ are given by: $(i = 1, \ldots, 4)$

$$\frac{\partial L_T}{\partial \alpha_i} = (1-\delta)\sum_{j=1}^{J_i} \frac{Y_{ij}\phi(u_{ij})}{\delta + (1-\delta)\Phi(u_{ij})} - \sum_{j=1}^{J_i} \frac{(n_{ij} - Y_{ij})\phi(u_{ij})}{1 - \Phi(u_{ij})},$$

$$\frac{\partial L_T}{\partial \beta_i} = (1-\delta)\sum_{j=1}^{J_i} \frac{Y_{ij}\phi(u_{ij})x_{ij}}{\delta + (1-\delta)\Phi(u_{ij})} - \sum_{j=1}^{J_i} \frac{(n_{ij} - Y_{ij})\phi(u_{ij})x_{ij}}{1 - \Phi(u_{ij})}.$$

It is easy to verify that

$$\mathbf{G}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\beta_1} & -\frac{\alpha_1}{\beta_1^2} & 0 & 0 & 0 & 0 & -\frac{1}{\beta_4} & \frac{\alpha_4}{\beta_4^2} \\ 0 & 0 & \frac{1}{\beta_2} & -\frac{\alpha_2}{\beta_2^2} & 0 & 0 & -\frac{1}{\beta_4} & \frac{\alpha_4}{\beta_4^2} \\ 0 & 0 & 0 & 0 & \frac{1}{\beta_3} & -\frac{\alpha_3}{\beta_3^2} & -\frac{1}{\beta_4} & \frac{\alpha_4}{\beta_4^2} \end{bmatrix},$$

and that the matrix $\mathbf{B}(\boldsymbol{\theta})$ is 8×8 and takes the form:

$$\mathbf{B}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{B}_1(\boldsymbol{\theta}_1) & 0 & 0 & 0 \\ 0 & \mathbf{B}_2(\boldsymbol{\theta}_2) & 0 & 0 \\ 0 & 0 & \mathbf{B}_3(\boldsymbol{\theta}_3) & 0 \\ 0 & 0 & 0 & \mathbf{B}_4(\boldsymbol{\theta}_4) \end{bmatrix},$$

where $\mathbf{B}_i(\boldsymbol{\theta}_i)$ is a 2×2 matrix whose elements are given in Appendix B, and 0 denotes a 2×2 matrix of zeroes.

The nointercept least-squares fit of the line through the points $\{(-\hat{\beta}_i, \hat{\alpha}_i)\}_{i=1}^4$ yields a slope of $\tilde{R} = 4.902$. Thus, the starting value for implementation of Henk's algorithm was $\tilde{\boldsymbol{\theta}}_0 = (-4.645, 0.948, -2.642, 0.539, -6.729, 1.373, -6.214, 1.268)'$. After 25 iterations the algorithm returns restricted maximum likelihood estimates that are identical to three decimal places, yielding $\tilde{\boldsymbol{\theta}} = (-1.946, 0.397, -0.156, 0.032, -6.916, 1.412, -3.490, 0.713)'$. The value of $L(\tilde{\boldsymbol{\theta}})$ is easily evaluated, and then it follows that $\Lambda = 138.45$. When compared to a chi-square distribution with 3 degrees of freedom, the p-value associated with the LRT is less than 0.001. The ability to test and formally reject the hypothesis of equal EC50 gives a practitioner more confidence to rank the toxins based on their estimated EC50 values.

4.2 QUADRATIC FORM TEST

Referring to Table 2, the MLEs $\hat{\boldsymbol{\theta}}_i$ $(i = 1, \ldots, 4)$ imply $\hat{\boldsymbol{\phi}} = (0.214, -0.925, -1.361)'$. Further calculations show that the estimated asymptotic variance–covariance matrices of $\hat{\boldsymbol{\theta}}_i$ $(i = 1, \ldots, 4)$ are:

$$S_1(\hat{\boldsymbol{\theta}}_1) = \begin{bmatrix} 0.284 & -0.0502 \\ -0.0502 & 0.00894 \end{bmatrix}, \qquad S_2(\hat{\boldsymbol{\theta}}_2) = \begin{bmatrix} 0.243 & -0.0429 \\ -0.0429 & 0.00763 \end{bmatrix},$$

$$S_3(\hat{\boldsymbol{\theta}}_3) = \begin{bmatrix} 0.263 & -0.0511 \\ -0.0511 & 0.0101 \end{bmatrix}, \qquad S_4(\hat{\boldsymbol{\theta}}_4) = \begin{bmatrix} 0.393 & -0.0949 \\ -0.0949 & 0.0232 \end{bmatrix},$$

and from (2.2) we find:

$$S(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} 3.173 \times 10^{-3} & 7.273 \times 10^{-5} & -8.232 \times 10^{-4} \\ 7.273 \times 10^{-5} & 2.105 \times 10^{-2} & 1.910 \times 10^{-2} \\ -8.232 \times 10^{-4} & 1.910 \times 10^{-2} & 7.373 \times 10^{-2} \end{bmatrix}.$$

It follows that $T_1 = \hat{\boldsymbol{\phi}}' S^{-1}(\hat{\boldsymbol{\theta}}) \hat{\boldsymbol{\phi}} = 59.29$. The trace terms in Equation (2.3), which represent the corrections relative to the underestimates provided by Equation (2.1) are, respectively, 4.017×10^{-5}, 9.20×10^{-5}, and 5.90×10^{-4}. Adding these terms to the diagonal of $S(\hat{\boldsymbol{\theta}})$ gives $S_c(\hat{\boldsymbol{\theta}})$, and a subsequent calculation gives $T_2 = \hat{\boldsymbol{\phi}}' S_c^{-1}(\hat{\boldsymbol{\theta}}) \hat{\boldsymbol{\phi}} = 58.95$.

Calibrated p-values associated with T_1 or T_2 were generated from the algorithm in Appendix A, as described in the previous section. The slope of the nointercept regression

of the points $(-\hat{\beta}_i, \hat{\alpha}_i)_{i=1}^4$ is $\tilde{R} = 4.9$, and this value together with the slopes $\{\hat{\beta}_i\}_{i=1}^4$, was used to generate independent bootstrap observations $\{Y_{ij}^* : 1 \le i \le 4, 1 \le j \le J_i\}$, where Y_{ij}^* has a binomial distribution with trial parameter n_{ij} and probability of success parameter $\delta + (1-\delta)\Phi[\hat{\beta}_i(x_{ij} - \tilde{R})]$ (recall from Section 1 that $\delta = 0.021$). The bootstrap calibrated p-values for both with T_1 or T_2 are both less than 0.001, implying the hypothesis of equal EC50 values for the four types of selenium is handily rejected.

5. POWER COMPARISON

The simulation study described in Section 3 was extended to compare the power of the proposed test statistics. Since calibration is critical when using T_1 or T_2, their uncalibrated versions were not evaluated in the power study. In addition, the simulation results confirm that the performance of T_1 and T_2 are very similar, so only the results for T_1 are reported. The design parameters for the power comparison coincide with those introduced in Section 3 for the size study, but here we explore alternatives to equal EC50 values.

It is easy to verify that $\phi_i = 0 \Leftrightarrow \beta_i\beta_1(R_i - R_1) = 0$ $(i = 2, 3, 4)$, and thus intuitively we can expect that the power of any test of H versus K will depend on both the slopes $\{\beta_i\}_{i=1}^4$ and $\{R_i\}_{i=1}^4$. Our size and power study reflects this observation in the following way. First, for the EC50 values we use four patterns of $\{R_i\}_{i=1}^4$. Letting R and $\Delta > 0$ denote arbitrary constants, the four patterns are (R, R, R, R), $(R, R, R, R+\Delta)$, $(R, R, R+\Delta, R+\Delta)$, and $(R, R+\Delta, R+\Delta, R+\Delta)$. The first of these patterns corresponds to the null hypothesis where all three ϕ_i $(i = 2, 3, 4)$ are zero. In the second pattern, only $\phi_2 = \phi_3 = 0$ and in the third pattern, only $\phi_2 = 0$. All three of the ϕ_i $(i = 2, 3, 4)$ are different from zero in the fourth pattern. Thus, the patterns get progressively farther away from the null in terms of Euclidean distance. For each of these four patterns, we also consider their symmetric counterparts that are obtained by replacing Δ by $-\Delta$. With $R = 5$ and $\delta = 0.021$ (as was chosen in Section 3 for the size study), we varied $\Delta \in \{0.2, 0.4\}$ to control the distance of the various alternatives from the null.

For the slopes, we again used the four equal sized intervals $B_1, \ldots, B_4$, and for each pattern of EC50 values and each slope interval B_j, we again selected 20 slope vectors $(\beta_1, \beta_2, \beta_3, \beta_4)'$ by randomly sampling four values from B_j. Then, for each combination of EC50 pattern and slope vector, we ran 1000 simulations to estimate the power of nominal 10% significance tests of H_0: equal EC50 values, based on both Λ and T_1, and using bootstrap calibrated p-values. The average power for the 20 repetitions corresponding to different slope vectors is reported as a summary of the power associated with the EC50 pattern and the slope interval B_j.

Table 4 reports the results for the power of the tests based on Λ and T_1. Both tests exhibit equivalent and good power for the alternatives that were considered. Comparing the power across all four slope intervals, it can be seen that the power increases as the magnitude of the slopes increases. This is an expected finding. (Additional simulations, not reported here, show that if the slopes are sampled across the four intervals, the power of the tests lies between the power obtained when the slopes are taken from within B_1 and B_4, respectively.) Also, as anticipated, the power increases as we move from $\Delta = 0.2$ to $\Delta = 0.4$. An

Table 4. Power of nominal 10% test of H_0: equal EC50 values.

		$\Delta = 0.2$		$\Delta = 0.4$	
EC50 pattern	Interval for slopes	Λ	T_1	Λ	T_1
$(R, R, R, R + \Delta)$	B_1	0.14	0.14	0.24	0.22
	B_2	0.21	0.21	0.47	0.49
	B_3	0.29	0.30	0.74	0.73
	B_4	0.37	0.38	0.86	0.88
$(R, R, R, R - \Delta)$	B_1	0.14	0.14	0.25	0.25
	B_2	0.21	0.21	0.52	0.52
	B_3	0.28	0.30	0.76	0.76
	B_4	0.37	0.40	0.88	0.90
$(R, R, R + \Delta, R + \Delta)$	B_1	0.16	0.14	0.32	0.25
	B_2	0.23	0.25	0.59	0.57
	B_3	0.35	0.37	0.82	0.85
	B_4	0.45	0.47	0.94	0.95
$(R, R, R - \Delta, R - \Delta)$	B_1	0.16	0.16	0.33	0.34
	B_2	0.25	0.26	0.64	0.66
	B_3	0.36	0.38	0.88	0.89
	B_4	0.46	0.48	0.95	0.96
$(R, R + \Delta, R + \Delta, R + \Delta)$	B_1	0.14	0.13	0.30	0.21
	B_2	0.21	0.21	0.51	0.46
	B_3	0.27	0.31	0.70	0.73
	B_4	0.37	0.38	0.84	0.86
$(R, R - \Delta, R - \Delta, R - \Delta)$	B_1	0.14	0.15	0.28	0.30
	B_2	0.21	0.23	0.56	0.58
	B_3	0.30	0.32	0.76	0.81
	B_4	0.37	0.39	0.88	0.90

interesting observation is that the $(R, R + \Delta, R + \Delta, R + \Delta)$ pattern does not have the highest power. Although it is the farthest away from the null pattern in terms of Euclidean distance, $(R, R, R + \Delta, R + \Delta)$ is the farthest away in terms of the more relevant $\boldsymbol{\phi}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}$ distance metric. The same observation is revealed when $-\Delta$ is used in the EC patterns. The dependence of power on a generalized distance is familiar from other contexts, including, for example, the power of the F-test in a completely randomized design where the power depends on a noncentrality parameter that is a function of a generalized distance.

6. EXTENSIONS AND SUMMARY

Although the LRT and quadratic form tests were derived in the context of wanting to test the equality of all of the EC50 values, the extension of the test procedures to a reduced subset of EC50 values (e.g., a particular pair) is straightforward. Similarly, extending the test procedures to consideration of other EC values and/or other forms of the binomial link function (e.g., logit) is equally straightforward. On the computational side, we note the algorithm we describe in Section 2.1 for obtaining the value of the numerator of the LRT is not the only way to proceed. Savvy users of the SAS package can utilize PROC NLIN to obtain that value. Readers interested in using this approach are referred to example 60.3 in

SAS (2008). Considering both the demonstrated computational tractability of the LRT and the diminished computational advantage of the quadratic form tests due to the necessity to calibrate their size, we recommend the LRT as a precursor test of equal EC50 values prior to doing a ranking analysis based on estimated EC50 values.

A. APPENDIX: BOOTSTRAP CALIBRATED p-VALUES

1. Initialize B and δ
2. Use $\{Y_{ij} : 1 \le i \le I, 1 \le j \le J_i\}$ to compute MLEs $(\hat{\alpha}_i, \hat{\beta}_i)_{i=1}^{I}$ and the observed test statistic T
3. Fit a no-intercept regression to the points $(-\hat{\beta}_i, \hat{\alpha}_i)_{i=1}^{I}$ and use the fitted slope $\tilde{R}$ to estimate the constrained EC50
4. For $k = 1$ to B
   For $i = 1$ to I
   For $j = 1$ to J_i
   Simulate Y_{ij}^* from a binomial distribution with trial parameter equal to n_{ij} and success probability $\delta + (1-\delta)\Phi[\hat{\beta}_i(x_{ij} - \tilde{R})]$
   Next j
   Next i
   Use the observations $\{Y_{ij}^* : 1 \le i \le I, 1 \le j \le J_i\}$ to compute bootstrap observation of the test statistic T_k^*
   Next k
5. Compute the calibrated p-value for the test statistic as the fraction of $\{T_k^*\}_{k=1}^{B}$ that exceed T.

B. APPENDIX: ELEMENTS FOR $\mathbf{B}_i(\boldsymbol{\theta}_i)$

$$\mathbf{B}_i(\boldsymbol{\theta}_i) = \begin{bmatrix} -b_{11}(\boldsymbol{\theta}_i) & -b_{12}(\boldsymbol{\theta}_i) \\ -b_{12}(\boldsymbol{\theta}_i) & -b_{22}(\boldsymbol{\theta}_i) \end{bmatrix}.$$

Recall that $u_{ij} = \alpha_i + \beta_i x_{ij}$,

$$\begin{aligned} b_{11}(\boldsymbol{\theta}_i) = (1-\delta) \sum_{j=1}^{J_i} n_{ij}[\delta + (1-\delta)\Phi(u_{ij})] \\ \times \left\{ \frac{[\delta + (1-\delta)\Phi(u_{ij})]\phi'(u_{ij}) - (1-\delta)\phi^2(u_{ij})}{[\delta + (1-\delta)\Phi(u_{ij})]^2} \right\} \\ - \sum_{j=1}^{J_i} n_{ij}(1-\delta)[1-\Phi(u_{ij})] \left\{ \frac{[1-\Phi(u_{ij})]\phi'(u_{ij}) + \phi^2(u_{ij})}{[1-\Phi(u_{ij})]^2} \right\}, \end{aligned}$$

$$b_{12}(\boldsymbol{\theta}_i) = (1-\delta)\sum_{j=1}^{J_i} n_{ij}[\delta + (1-\delta)\Phi(u_{ij})]x_{ij}$$
$$\times \left\{ \frac{[\delta + (1-\delta)\Phi(u_{ij})]\phi'(u_{ij}) - (1-\delta)\phi^2(u_{ij})}{[\delta + (1-\delta)\Phi(u_{ij})]^2} \right\}$$
$$- \sum_{j=1}^{J_i} n_{ij}(1-\delta)[1-\Phi(u_{ij})]x_{ij} \left\{ \frac{[1-\Phi(u_{ij})]\phi'(u_{ij}) + \phi^2(u_{ij})}{[1-\Phi(u_{ij})]^2} \right\},$$

$$b_{22}(\boldsymbol{\theta}_i) = (1-\delta)\sum_{j=1}^{J_i} n_{ij}[\delta + (1-\delta)\Phi(u_{ij})]x_{ij}^2$$
$$\times \left\{ \frac{[\delta + (1-\delta)\Phi(u_{ij})]\phi'(u_{ij}) - (1-\delta)\phi^2(u_{ij})}{[\delta + (1-\delta)\Phi(u_{ij})]^2} \right\}$$
$$- \sum_{j=1}^{J_i} n_{ij}(1-\delta)[1-\Phi(u_{ij})]x_{ij}^2 \left\{ \frac{[1-\Phi(u_{ij})]\phi'(u_{ij}) + \phi^2(u_{ij})}{[1-\Phi(u_{ij})]^2} \right\}.$$

SUPPLEMENTAL MATERIALS

R-code: R-code for computing the LRT for the Selenium data example is available from the *JABES* website. (R Code for LRT With Selenium Data.txt)

ACKNOWLEDGMENTS

The authors thank the editor and two reviewers for constructive comments that led to improvements in our development and presentation of the LRT.

[Received September 2007. Revised October 2008.]

REFERENCES

Abot, A. R., Moscardi, F., Fuxa, J. R., Sosa-Gomez, D. R., and Richter, A. R. (1995), "Susceptibility of Populations of *Anticarsia gemmatalis* (Lepidoptera: Noctuidae) From Brazil and the United States to a Nuclear Polyhedrosis Virus," *Journal of Entomological Science*, 30, 62–69.

Brown, A. W. A. (1958), *Insecticide Resistance in Anthropods*, Geneva: World Health Organization.

Chernick, M. R. (1999), *Bootstrap Methods: A Practitioners Guide*, New York: Wiley.

Cox, C. (1990), "Fieller's Theorem, the Likelihood and the Delta Method," *Biometrics*, 46, 709–718.

Henk, D. F. J. (1985), "The Use of Generalized Inversed in Restricted Maximum Likelihood," *Linear Algebra and Its Applications*, 70, 225–240.

Liu, T. X., Hutchison, W. D., Chen, W., and Burkness, E. C. (2003), "Comparative Susceptibilities of Diamondback Moth (Lepidoptera: Plutellidae) and Cabbage Looper (Lepidoptera: Noctuidae) From Minnesota and South Texas λ-Cyhalothrin and Indoxacarb," *Journal of Horticultural Entomology*, 94, 1230–1236.

Payton, M. E., Greenstone, H. H., and Schenker, N. (2003), "Overlapping Confidence Intervals or Standard Error Intervals: What Do They Mean in Terms of Statistical Significance?" *Journal of Insect Science*, 3, 34–39.

SAS (2008), *SAS/STAT User's Guide Version 9.2*, Cary, NC: SAS Institute Inc.

Schenker, N., and Gentleman, J. F. (2001), "On Judging the Significance of Differences by Examining Overlap Between Confidence Intervals," *The American Statistician*, 55, 182–186.

Sokal, R. R., and Rohlf, F. J. (1969), *Biometry*, San Francisco: Freeman.

Van Birgelen, A. P., Van der Kolk, J., Fase, K. M., Bol, I., Poiger, H., Brouwer, A., and Van den Berg, M. (1995), "Subchronic Dose-Response Study of 2,3,7,8-Tetrachlorodibenzo-*p*-Dioxin in Female Sprague–Dawley Rates," *Toxicology and Applied Pharmacology*, 132, 1–13.

Van Gestel, C. A. M., and Hensbergen, P. J. (1997), "Interaction of Cd and Zn Toxicity for *Folsomia candida* Willem (Collembola: Isotomidae) in Relation to Bioavailability in Soil," *Environmental Toxicology and Chemistry*, 16, 1177–1186.

Zieger, M., Haag, R., Hockel, J., Schrenk, D., and Schmitz, H.-J. (2001), "Inducing Effects of Dioxin-Like Polychlorinated Biphenyls on CYP1A in the Human Hepatoblastoma Cell Line HepG2, the Rat Hepatoma Cell Line H4IIE, and Rat Primary Hepatocytes: Comparison of Relative Potencies," *Toxicological Sciences*, 63, 65–73.