SETAC PRESS

Environmental Toxicology and Chemistry, Vol. 19, No. 1, pp. 169–174, 2000
Printed in the USA
0730-7268/00 $9.00 + .00

Annual Review

BIOEQUIVALENCE APPROACH FOR WHOLE EFFLUENT TOXICITY TESTING

RAKESH SHUKLA,*† QIN WANG,‡ FLORENCE FULK,§ CHUNQIN DENG,† and DEBRA DENTON‖
†Department of Environmental Health, University of Cincinnati Medical Center, Cincinnati, Ohio 45267, USA
‡Biostatistics and Epidemiology, Abt Associates Clinical Trials, 55 Wheeler Street, Cambridge, Massachusetts 02138, USA
§U.S. Environmental Protection Agency, National Exposure Research Laboratory, Cincinnati, Ohio 45268
‖U.S. Environmental Protection Agency, Region 9, San Francisco, California 94105

(*Received* 28 *January* 1999; *Accepted* 1 *July* 1999)

Abstract—Increased use of whole effluent toxicity (WET) tests in the regulatory arena has brought increased concern over the statistical analysis of WET test data and the determination of toxicity. One concern is the issue of statistical power. A number of WET tests may pass the current hypothesis test approach because they lack statistical power to detect relevant toxic effects because of large within-test variability. Additionally, a number of WET tests may fail the current approach because they possess excessive statistical power, as a result of small within-test variability, and detect small differences that may not be biologically relevant. The strengths and limitations of both the traditional hypothesis test approach and the bioequivalence approach for use in the National Pollutant Discharge Elimination System program were evaluated. Data from 5,213 single-concentration, short-term chronic WET tests with *Ceriodaphnia dubia* provided the database for analysis. Comparison of results between the current approach and the bioequivalence approach indicates that the current approach to WET testing is generally sound but that adopting the proposed bioequivalence approach resolves concerns of statistical power. Specifically, within this data set, applying the bioequivalence approach resulted in failure for tests with relatively large test variability and a pass for tests with relatively small within-test variability.

Keywords—Toxicity testing Bioequivalence testing Hypothesis testing Practically equivalent toxicity

INTRODUCTION

Whole effluent toxicity (WET) tests are accepted standard methods by the U.S. Environmental Protection Agency in the assessment of the toxicity of effluents. Currently, there are approx. 6,500 permits requiring WET tests for compliance monitoring in the National Pollutant Discharge Elimination System program. One of the goals of WET tests, within the framework of National Pollutant Discharge Elimination System permits, is to determine the safe or no-adverse-effect concentration of the effluent or receiving water. The U.S. Environmental Protection Agency currently recommends two statistical approaches to determine a safe concentration for each biological endpoint (e.g., survival, growth, or reproduction) [1–3]. One approach is derivation of the no-observed-effect concentration by a hypothesis test that equates biological significance with statistical significance. The second approach is estimation of an inhibition concentration or effective concentration endpoint that reduces the control response by 25%. The expanded use of WET tests in the National Pollutant Discharge Elimination System program has brought increased attention to the statistical analysis of toxicity test data.

In September 1995, a Society of Environmental Toxicology and Chemistry Pellston workshop was convened to discuss unresolved scientific issues and to highlight significant research needs associated with WET testing. Workshop participants were WET testing national experts representing government, academia, and industry. One recommendation of the Pellston workshop [4] was to "immediately instigate studies to evaluate improvements in the statistical analysis of WET test data. These studies should include, but not necessarily be limited to, the following activities: (a) investigate the implications of concurrent application of No Observed Effect Concentration/Minimum Significant Difference, tests of bioequivalence, and effective concentration estimators." In response to this recommendation, a study to evaluate the utility of the bioequivalence approach for the WET program was initiated. The strengths and limitations of both the traditional hypothesis test approach and the bioequivalence approach in the interpretation of test results for use in the National Pollutant Discharge Elimination System program were evaluated.

Current approach

Here we test the null hypothesis, H_0 ($\mu_c - \mu_t = 0$), where μ_c and μ_t refer to the population means for control and effluent groups, respectively, against the alternative hypothesis, H_a ($\mu_c - \mu_t > 0$) or ($\mu_c - \mu_t < 0$), which will depend on the biological response of interest (i.e., mortality, growth, or reproduction). We use the following test statistic:

$$t = \frac{\bar{X}_c - \bar{X}_t}{\sqrt{S_p^2 \cdot \left(\frac{1}{n_c} + \frac{1}{n_t}\right)}}$$

where

$\bar{X}_c$ = mean for the control

$\bar{X}_t$ = mean for the effluent concentration

$$S_p = \sqrt{\frac{(n_t - 1)S_t^2 + (n_c - 1)S_c^2}{n_t + n_c - 2}}$$

S_c^2 = variance for the control, S_t^2 = variance for the effluent

* To whom correspondence may be addressed (rakesh.shukla@uc.edu).

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Approval does not signify that the contents necessarily reflect the views or policies of the agency nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

concentration, n_c = number of replicates for the control, n_t = number of replicates for the effluent concentration.

If the calculated t is greater than a Student's t distribution at the level of significance with $n_t + n_c - 2$ degrees of freedom, effluent concentration is declared 'toxic.' Otherwise, the effluent is considered nontoxic. Appropriate adjustments can be made for non-normal data or non-homogenous variances.

The type I or false-positive error (α) in the current hypothesis test approach is the risk of concluding an effluent is toxic when, in fact, it is not; it is typically set at 0.01 or 0.05. Operational consequences of committing a type I error are of concern, because it results in failing a WET test and potentially a permit limit. This may lead to more frequent testing and to additional monetary consequences to the regulated community. The type II or false-negative error (β) is the risk of concluding the effluent is not toxic when it is, in fact, toxic. The type II error is neither fixed nor controlled at a known level in the current hypothesis test approach. Consequences of committing a type II error are also of concern, because it results in the continued discharge of a toxic effluent. Regulators, in an attempt to control the type II error rate, specify such test design elements as the number of replicates, test acceptability criteria, minimum significant difference criterion, the test statistic applied, and the type I error rate, among others. It is straightforward to specify and control type I error under the current hypothesis test approach. However, the specification and control of type II error (and thus the power) is more complex, thus being a major cause of concern with the current hypothesis test approach. It can be expressed as follows: A somewhat larger-than-expected portion of WET tests may pass the current approach because of large within-test variability and thus may lack the expected statistical power to detect relevant toxic effects, and a somewhat larger-than-expected portion of WET tests may fail the current approach because of small within-test variability and thus may possess the statistical power to detect small effects that may not be biologically relevant [5].

Proposed bioequivalence approach

It is improbable that the population mean responses of the effluent and control are exactly the same. They may differ by such a small amount that even if statistically significant, it would not be considered ecologically significant. A more relevant approach may be to rephrase the hypothesis from "Are the mean responses the same?" to "Do the mean responses differ by more than some amount?" An interval of practically equivalent responses could be initially determined and tested as an interval hypothesis. This approach is known as a test of bioequivalence. The current hypothesis test approach and the bioequivalence approach have two major differences. First, in the bioequivalence approach, the null and the alternative hypotheses are interchanged. The null hypothesis becomes a statement of difference between the control response and the effluent concentration response ($\mu_c - \mu_t \geq \theta_0$ or $\mu_c - \mu_t \leq \theta_0$). Second, the amount of difference is not greater than or less than zero but greater than or less than a specified amount, θ_0. It is the replacement of a zero difference with an acceptably small nonzero difference, θ_0, that provides the theoretical basis for interchanging the hypotheses. Simply switching the hypotheses often leads to contradictory conclusions for a given experiment [6]. Type I and II errors in the bioequivalence approach are reversed; they represent type II and I errors, respectively, of the current approach.

Although tests of bioequivalence have gained acceptability in drug development and evaluation, they have not gained acceptance in toxicity testing in the environmental regulatory arena. Erickson and McDonald [7], proposed bioequivalence testing as an alternative to the current hypothesis test approach in WET testing. In their article, the bioequivalence test is constructed as a test of a proportionality constant expressed as the ratio of the mean control response to the mean effluent concentration response. The bioequivalence approach was compared with the current approach for both single-concentration and multiple-concentration toxicity tests for a survival endpoint. They also assessed the effect of increasing the number of replicates on the agreement of conclusions reached by both approaches.

The bioequivalence testing problem can be formulated as a ratio hypothesis or, alternatively, as a difference hypothesis. It is customary to use log-normal models in bioequivalence studies of drug development, but, sometimes, a normal model for the data is more appropriate. Should one use the ratio hypothesis or the difference hypothesis? If the data appear to be clearly normal or clearly log normal, then the choice of an appropriate hypothesis is clear. However, the small sample sizes typically used in WET tests will produce tests of normality that have fairly low power in either the original or log scale. This may lead to an equivocal verdict regarding the distributional property of the data. In such situations, either the ratio or difference hypothesis can be used. In the present study, a comparison of the bioequivalence approach to the current approach is made by a practically equivalent toxicity (PET) level, where we use a difference hypothesis instead of a ratio hypothesis.

The PET level is the difference, θ_0, between the mean control response and the mean effluent concentration response that is considered to be an acceptable level of effect. To illustrate a test of bioequivalence with a PET level, consider the example of comparing the mean growth response in the control to the mean growth response in the effluent concentration. In the bioequivalence context, the null and alternative hypotheses are stated as follows:

$$H_o: \mu_c - \mu_t \geq \theta_0$$

$$H_a: \mu_c - \mu_t < \theta_0$$

The test statistic for the test of bioequivalence is

$$t = \frac{\theta_0 - \bar{X}_c - \bar{X}_t}{\sqrt{S_p^2 \cdot \left(\frac{1}{n_c} + \frac{1}{n_t}\right)}}$$

where θ_0 is the PET level and all other parameters are as previously defined. The difference in the mean growth response in the control and effluent concentration is significantly less than θ_0 if the calculated t is greater than a Student's t distribution at the level of significance with $n_t + n_c - 2$ degrees of freedom (i.e., the effluent is not toxic). Appropriate adjustments can be made for nonnormal data or nonhomogenous variances between the groups.

The intent of this study was to evaluate the utility of the bioequivalence procedure in the determination of the toxicity of effluents and receiving waters in comparison to the current hypothesis testing approach.

MATERIALS AND METHODS

Toxicity tests

Reproductive data from 5,213 short-term chronic WET tests with *Ceriodaphnia dubia* were used in the comparative anal-

ysis. The data were collected by the North Carolina Department of Environment, Health, and Natural Resources, Division of Environmental Management, as part of their National Pollutant Discharge Elimination System program. Each test consists of a single effluent concentration tested against a control, with 12 replicates per test group. This single-concentration test is typically referred to as a pass/fail test. A failure occurs when a statistically significant difference between the control response and the effluent response is detected. Similarly, a pass is declared when no such statistical difference is detected. Following the current hypothesis test approach, statistically significant differences ($p < 0.01$) were detected in mean reproduction in 750 of the 5,213 tests (14% failures). Whole effluent toxicity tests that did not meet the test acceptance criteria, 80% control survival and an average of 15 young per surviving female in the control, were excluded from the data set.

Selection of PET level

To compare the test results by the current approach with results by the bioequivalence approach, it is necessary to first specify the PET level (θ_0). A satisfactory level can be established in several ways. One reasonable approach is to hold constant the total number of tests that passed or failed by either method and establish the PET level such that the discordant pairs of tests is an evenly balanced distribution. A discordant result is one in which the current approach resulted in a pass and the bioequivalence approach resulted in a failure or in which the current approach resulted in a failure and the bioequivalence approach resulted in a pass. Specification of the PET level in this manner maintains the same frequency of passed or failed tests by either method while reducing both the number of false-positive and false-negative results. Increasing or decreasing WET test compliance rates may engender resistance from both dischargers and regulators. A bioequivalence approach that only reduces statistical errors, without changing the number of WET tests that pass or fail overall, may prove to be more acceptable. A PET level set in this manner, unfortunately, does not address the question of what is ecologically significant, but it can begin the dialogue to address what is considered an acceptable level of ecological effect.

Determination of the PET level, θ_0, was empirically derived by trial and error. The total number of tests that passed and failed was first determined by applying the current approach to the data set with a type I error of 0.01. The bioequivalence approach was iteratively applied to the same data set, also with a type I error of 0.01, to determine a PET level that resulted in the same number of passed and failed tests as the current approach and balanced the distribution of discordant test results.

RESULTS

To characterize the population of the 5,213 WET tests, means, standard deviations (SDs), and maximum and minimum values were tabulated for all tests for the control mean, control SD, pooled SD for control and effluent, and the difference in mean control response and mean effluent response. Summary statistics were also calculated for the control SD and the difference in mean control response and mean effluent response for each subset of passed and failed tests, where passing or failure was determined by the current hypothesis

Table 1. Distribution of whole effluent toxicity tests cross-classified by current approach versus bioequivalence approach

	Bioequivalence approach[a]		
Current approach	Passed tests (nontoxic)	Failed tests (toxic)	Total
Passed tests (nontoxic)	4,256	207	4,463
Failed tests (toxic)	207	543	750
Total	4,463	750	5,213

[a] At practically equivalent toxicity level of 11.59.

young produced. Each WET test consisted of 12 replicates per group; the testing was performed at a significance level of 0.01. After determination of the PET level, the 5,213 WET tests were analyzed using the bioequivalence approach. In all cases, a failure is a determination of toxicity, and a pass is a determination of no toxicity.

Summary statistics of all WET tests included in the comparative analysis were obtained. Although the average difference between the mean control response and the mean effluent response is small (0.56), the range of the difference is wide (−33–34). This is consistent with the range of the average number of young in the control group (15–44), indicating some level of inherent variability in the response variable. The average variability in the control group and the average pooled SD are similar (4.54 vs 4.75). However, there is a wide range in the variability estimates (0.79–15.17 for the control group and 0.97–15.37 for the pooled estimate). The between-test differences are reflected in the range of the variability estimates.

A comparison of the descriptive statistics of the passed tests to the failed tests indicates that the average difference between the mean control response and the mean effluent response is almost zero for the passed tests. The average difference for the failed tests is 11.42. Some tests passed with a difference in means as high as 9.25, and some tests failed with a difference in means as low as 2.08. However, there is almost no difference in the mean control group SD between passed and failed tests.

Results of the determination of an optimum PET level are presented in Table 1. For $\theta_0 = 11.59$, the two sets of discordant pairs were balanced at 207 tests. This choice of θ_0 maintains the overall ratio of passed versus failed tests at the same level for both statistical approaches and does not bias the overall conclusions in favor of one statistical approach. Additionally, at this PET level, 92% of the tests reached the same conclusion with the current approach and the bioequivalence approach.

Concordant results

Summary statistics for the concordant and discordant results are given in Table 2. The average difference in mean control response and mean effluent response for tests that passed by both approaches is near zero. For tests that failed by both approaches, the mean difference is almost 14. The range in mean differences for the concordant tests does not overlap; the minimum difference in mean responses for failed tests (5.92) is greater than the maximum difference in mean responses for tests that passed (5.75). The average pooled SD and the range of the pooled SDs are similar between the two sets of concordant tests.

Table 2. Summary statistics of whole effluent toxicity test cross-classified by current approach versus bioequivalence approach, short-term *Ceriodaphnia dubia* reproductive data

Current approach		Bioequivalence approach[a]					
		Passed tests (nontoxic)			Failed tests (toxic)		
		Mean	Minimum	Maximum	Mean	Minimum	Maximum
Passed tests (nontoxic)	Control − effluent[b]	−1.57	−33.17	5.75	5.08	−1.08	9.25
	Pooled SD	4.27	0.97	13.03	8.30	5.70	15.37
	Estimated post hoc power[c]	99.99	39	100	81	28	99
Failed tests (toxic)	Control − effluent[b]	5.15	2.08	9.25	13.82	5.92	34.08
	Pooled SD	3.26	1.59	5.41	4.88	1.14	13.85
	Estimated post hoc power[c]	>99.99	99.5	100	99.9	34	100

[a] At practically equivalent toxicity level of 11.59.
[b] Difference in mean response (control − effluent).
[c] Estimated post hoc power at $\mu_c - \mu_t = 11.59$; mean represented by median.

by design, the type II error at $\theta_0 = 11.59$ should also be around 1%. Thus, the expected power is approximately 99%. Results indicate that >90% of the concordant tests had a power >99% to detect a difference in mean response of at least 11.59. A few had power as low as 34 or 39%. The median power value was used to represent the mean in Table 2 because of the skewed distribution of the power values.

Discordant results

The two sets of discordant tests have similar distributions of the difference in mean response, both from passed-to-failed tests and failed-to-passed tests. The most obvious difference in the two types of discordant pairs is the level of the pooled SD. For those tests that passed by the current hypothesis test approach and failed by the bioequivalence approach (passed-to-failed tests), the mean pooled SD was more than twice the same measure for the failed-to-passed tests. Also, the minimum pooled standard deviation (5.70) for the passed-to-failed tests was larger than the maximum pooled SD estimated for the failed-to-passed tests (5.41). These results indicate why the bioequivalence approach is a win–win solution for WET testing. From the regulator's perspective, there are WET tests that tend to pass the current approach simply because of large variability in the tests results because they do not possess adequate statistical power to detect relevant toxic effects from the effluent exposure. Notice that the maximum difference among tests that passed by both approaches is about the same magnitude as the average difference among those tests that passed in the current approach and failed in the bioequivalence approach. Table 3 lists the tests with the 10 smallest and 10 largest differences for the passed-to-failed tests. Although the smallest difference in means for passed-to-failed tests is −1.08, such a test does not pass the bioequivalence approach simply because the pooled SD (12.74) for this test was three times the average pooled SD for tests that passed by both approaches (4.27). Overall, the tests with the smallest differences in means were associated with relatively large pooled SDs. These tests have low power to detect toxicity at the PET level of 11.59. Even among those tests that had larger mean differences, the power was relatively low compared with the power in other tests.

Table 3. Extreme differences associated with whole effluent toxicity tests that passed the current approach and failed the bioequivalence approach, short-term *Ceriodaphnia dubia* reproductive data

Difference in means	Pooled SD	Estimated power at $\theta_0 = 11.59$
Smallest difference in means		
−1.08	12.74	41
0.16	12.02	45
0.25	12.67	41
1.25	11.59	49
1.58	9.94	64
1.67	10.82	55
1.67	10.71	56
1.83	10.25	61
1.92	9.67	67
1.92	9.59	67
Largest difference in means		
9.25	9.33	70
9.17	9.37	70
9.00	9.21	71
9.00	9.18	72
8.75	8.75	76
8.67	9.13	72
8.50	9.10	73
8.50	11.19	52
8.42	10.26	61

From the discharger's perspective, there are tests that failed in the current approach but passed in the proposed bioequivalence approach. The single most important reason for these tests to fail was the small pooled SD, which gives them substantially high statistical power. Having more than the necessary level of power in a given test, in the current approach, may be detrimental to the discharger by resulting in regulatory action. The situation is automatically rectified in the proposed bioequivalence approach because tests with more than adequate power tend to pass and result in a determination of no toxicity. Thus, there is a natural incentive for the regulated community to achieve lower levels of variability by conducting tests with larger numbers of replicates, maintaining high quality assurance/quality control standards, and by seeking out

Table 4. Extreme differences associated with whole effluent toxicity tests that failed the current approach and passed the bioequivalence approach, short-term *Ceriodaphnia dubia* reproductive data

Difference in means	Pooled SD[a]	Estimated power at $\theta_0 = 11.59$
Smallest difference in means		
2.84	1.59	100
2.17	1.89	100
2.67	2.32	100
2.67	2.44	100
2.75	2.25	100
2.75	2.24	100
2.92	2.80	100
2.92	2.39	100
3.00	2.76	100
3.00	2.21	100
Largest difference in means		
9.25	2.26	100
9.17	2.01	100
9.08	2.32	100
8.00	2.83	100
8.00	2.89	100
7.91	3.09	100
7.91	2.50	100
7.75	3.08	100
7.75	2.86	100
7.58	3.84	100

is absent in the current hypothesis test approach. Table 4 lists the failed-to-passed tests with the 10 smallest differences and the 10 largest differences. The lower half of the table clearly indicates that smaller pooled SDs allow larger differences in means to pass by the bioequivalence approach when they would have failed in the current approach.

DISCUSSION

With the increased use of WET tests in the regulatory arena has come increased concern over the statistical analysis of WET test data and the determination of toxicity. These concerns are expressed by both the discharger and the regulating authority. One such concern revolves around the issue of power. A portion of WET tests pass by the current approach and lack the statistical power to detect relevant toxic effects because of large within-test variability. Additionally, a portion of WET tests fail by the current approach and possess the statistical power to detect small differences that may not be biologically relevant because of small within-test variability. Comparison of results between the current hypothesis test approach and the proposed bioequivalence approach indicates that the current approach to WET testing is generally sound. However, the results also indicate that by adopting the proposed bioequivalence approach, the positive features of the current approach are maintained while incorporating the benefits of the bioequivalence approach. Specifically, within this data set, applying the bioequivalence approach resulted in failure for tests with large test variability and in a pass for tests with small within-test variability. Thus, the bioequivalence approach addresses the two major concerns expressed by both parties and results in a win–win solution for the WET testing program.

Several issues need to be mentioned. Although an upper limit of a practically equivalent toxicity level for the effluent response exists implicitly in the current approach, such a choice needs to be made explicitly in the bioequivalence approach. The choice of a PET level is a biological, statistical, social, and regulatory issue. In the current study, the PET level was determined empirically on the basis of maintaining the same number of passed and failed tests by both approaches while balancing the number of discordant results. A method that incorporates the response variability in the control group provides a more theoretical basis for the selection of θ_0. The following equation relates the PET level, type I and type II errors, common estimates of within-test error variance, and the number of replicates per group [8]:

$$\theta_0 = [(t_{(1-\alpha)(2n-2)} + t_{(1-\beta)(2n-2)}) \cdot S_p\sqrt{2/n}]$$

If the type I and type II errors are to be the same, then the above equation reduces to

$$\theta_0 = 2(t_{(1-\alpha)(2n-2)}) \cdot S_p\sqrt{2/n} = K \cdot S_p$$

where

$$K = 2(t_{(1-\alpha)(2n-2)}) \cdot \sqrt{2/n}$$

Table 5. Sample size and/or practically equivalent toxicity level determination[a]

Sample size (N per group)	K $\alpha = \beta = 0.05$	K $\alpha = \beta = 0.01$
2	5.84	13.92
3	3.48	6.12
4	2.75	4.44
5	2.40	3.66
6	2.09	3.19
7	1.90	2.86
8	1.76	2.62
9	1.65	2.43
10	1.55	2.28
11	1.47	2.15
12	1.40	2.05
13	1.34	1.95
14	1.29	1.87
15	1.24	1.80
20	1.06	1.59

[a] For a given S_p, $\theta_0 = K \cdot S_p$.

One can obtain the required sample size, n, per group for a given θ_0, or determine θ_0, for any given choice of n. As an example, assume that an estimate of the population variance for the response in question is 25 (i.e., $S_p = 5$) and $\alpha = \beta = 0.05$. Table 5 can be used to obtain K for a given sample size per group. If $n = 10$, then $K = 1.55$ and $\theta_0 = 1.55 \cdot 5 = 7.75$ number of young. Thus, a PET level of 7.75 can be used with 10 replicates per group assuming a pooled estimate of within-test variability of 5. Under these specifications, the type I error in the bioequivalence approach will be at most 5% when the true mean difference between the control and effluent response is 7.75. Note that at these specifications the type II error is also 5% when the control and effluent have the same response ($\theta_0 = 0$). In other words, if the true responses for the control and the effluent are the same, then we have at least 95% power to declare that the effluent is nontoxic using the proposed bioequivalence approach. It is worthwhile to point out that the PET level was arrived at in the present data analysis by trial and error in an attempt to minimize the discordant pairs as well as to keep the total number of passed and failed tests unchanged, thus *not* biasing the conclusion in favor of either approach.

A note of caution is, however, necessary at this point. When the PET level is determined on the basis of a difference between the mean response in the control and the mean response

in the effluent, then relatively large reductions in response relative to the control may pass when the control response itself is low. For example, the test acceptability criteria for control performance in the *C. dubia* 7-d chronic test is an average of 15 young per surviving female. Table 2 lists 9.25 young as the maximum difference between the control and effluent response that passed with the bioequivalence approach with $\theta_0 = 11.59$. A difference of 9.25 represents a 62% reduction in reproduction if the control mean performance was only 15 young. Although such situations like this may be infrequent in practice, raising control performance criteria may be necessary if the proposed bioequivalence approach is adopted.

In general, adoption of the bioequivalence approach is highly dependent on the selection of a reasonable PET level. The PET level, θ_0, should represent a practical biological effect, be a constant value, and be decided before the WET test is conducted. Historical data are an important resource for choosing an acceptable PET level. Further research related to bioequivalence testing in multiple-concentration WET tests is necessary before it can be considered as the statistical analysis of choice in the WET testing arena. The replication of the findings of this study with additional data sets, including other species, will strengthen the case for its use.

Acknowledgement—This work was supported in part by cooperative agreement EPA CR820480 between the University of Cincinnati and the U.S. Environmental Protection Agency. We express our sincere thanks to Larry Ausley for providing us with the data. We also thank Randall Marshall for valuable review of earlier versions of the manuscript.

REFERENCES

1. Lewis PA, Klemm DJ, Lazorchak JM, Norberg-King TJ, Peltier WH, Heber MA. 1994. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to freshwater organisms, 3rd ed. EPA600/4-91/002. U.S. Environmental Protection Agency, Cincinnati, OH.
2. Klemm DJ, Morrison GE, Norberg-King TJ, Peltier WH, Heber MA. 1994. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to marine and estuarine organisms, 2nd ed. EPA600/4-91/003. U.S. Environmental Protection Agency, Cincinnati, OH.
3. Chapman GA, Denton DL, Lazorchak JM. 1995. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to west coast marine and estuarine organisms, 2nd ed. EPA/600/R-95-136. U.S. Environmental Protection Agency, Cincinnati, OH.
4. Chapman GA, et al. 1996. Methods and appropriate endpoints. In Grothe DR, Dickson KL, Reed-Judkins DK, eds, *Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving System Impacts*. Society of Environmental Toxicology and Chemistry, Pensacola, FL, USA, pp 97–99.
5. Denton DL, Norberg-King TJ. 1996. Whole effluent toxicity statistics: A regulatory perspective in whole effluent toxicity testing: An evaluation of methods and prediction of receiving system impacts. In Grothe DR, Dickson KL, Reed-Judkins DK, eds, *Whole Effluent Toxicity Testing: An Evaluation of Methods and Prediction of Receiving System Impacts*. Society of Environmental Toxicology and Chemistry, Pensacola, FL, USA, pp 83–102.
6. Guenther WC. 1965. *Concepts of Statistical Inference*, 2nd ed. McGraw-Hill, New York, NY, USA, pp 217–218.
7. Erickson WP, McDonald LL. 1995. Tests for bioequivalence of control media and test media in studies of toxicity. *Environ Toxicol Chem* 14:1247–1258.
8. Cochran WG, Cox GM. 1992. *Experimental Design*, 2nd ed. John Wiley & Sons, New York, NY, USA.