SETAC – Brazil

J. Braz. Soc. Ecotoxicol., v. 2, n. 2, 2007, 139-145

JBSE

Selection of Relevant Effect Levels for Using Bioequivalence Hypothesis Testing

E. Bertoletti,* S. V. Buratini, V. A. Prósperi, R. P. A. Araújo & L. I. Werner

Companhia de Tecnologia e Saneamento Ambiental – CETESB/SP, Setor de Ecotoxicologia Aquática, Av. Prof. Frederico Hermann Jr., 345, CEP 05459-900, São Paulo, SP

(Received December 18, 2006; Accepted April 25, 2007)

ABSTRACT

The use of classic hypotheses tests based on a null hypothesis of equal means frequently promote the occurrence of false positives and false negatives in ecotoxicological assays results. The use of criteria or appropriate statistical analyses is recommended to prevent these occurences and to guarantee the quality of the results from the biological as well statistical point of view. Therefore, the relevant effect levels were established for ecotoxicological assays with *Daphnia similis*, *Ceriodaphnia dubia*, *Danio rerio, Hyalella azteca, H. meinerti*, *Lytechinus variegatus* and *Mysidopsis juniae*. Such effect levels were estimated on the basis of the 75th percentile of the Minimum Significant Difference (MSD) of a historical series of analytical results. The estimated values were used to evaluate the allowable variability of the analytical results. It was demonstrated that the integration between the relevant effect levels and the bioequivalence hypothesis testing minimizes the occurrence of false positive results relative to those observed using traditional hypothesis testing.

Key words: analytical variability, ecotoxicological assay, minimum significant difference, relevant effect level.

RESUMO

Níveis de efeito relevante para o uso em testes de hipóteses

O uso de testes de hipóteses clássicos, baseados na hipótese nula de igualdade entre as médias, freqüentemente provoca falsos positivos e falsos negativos nos resultados de ensaios ecotoxicológicos. Para evitar essas ocorrências é recomendável a utilização de critérios ou análises estatísticas apropriadas, com vistas a garantir a qualidade dos resultados tanto do ponto de vista biológico como do estatístico. Considerando esses aspectos foram estabelecidos os níveis de efeito relevante para ensaios ecotoxicológicos com *Daphnia similis, Ceriodaphnia dubia, Danio rerio, Hyalella azteca, H. meinerti, Lytechinus variegatus* e *Mysidopsis juniae*. Tais níveis de efeito foram estimados com base no 75º percentil da diferença mínima significativa (DMS) de uma série histórica de resultados analíticos. Os valores estimados mostraram-se úteis para avaliar a variabilidade admissível dos resultados analíticos. Foi demonstrado que a integração entre os níveis de efeito relevante e o teste de hipóteses por bioequivalência minimiza a ocorrência de falsos positivos observados nos resultados calculados pelo teste de hipótese tradicional.

Palavras-chave: diferença mínima significativa, nível de efeito relevante, ensaio ecotoxicológico, variabilidade analítica.

INTRODUCTION

Hypotheses tests using the null hypothesis of equal means have been used since the 1960s to analyze the results of ecotoxicological assays with aquatic organisms. However, some researchers (Erickson & McDonald, 1995; Chapman *et al*., 1996; Garrett, 1997; Shukla *et al*., 2000; Buratini & Bertoletti, 2006) have demonstrated that these statistical tools oftentimes are inappropriate due to the occurrence of false positives (the effects are statistically but not biologically significant) or false negatives (adverse effects exist, but are not detected by statistical analysis).

*Corresponding author: Eduardo Bertoletti, e-mail: eduardob@cetesbnet.sp.gov.br.

Erickson & McDonald (1995) proposed the use of bioequivalence hypothesis testing, more accurately called equivalence testing, to avoid erroneous conclusions based on statistical analysis of ecotoxicological data. The critical point of the bioequivalence approach is the establishment of the adverse effect level that can be considered relevant relative to the experimental control.

Some studies (Erickson & McDonald, 1995; Shukla *et al.*, 2000) have suggested that the relevant effect levels depend on a given laboratory's performance, particularly regarding the variance of the experimental control, number of replicates and power of the statistical analysis. Other studies (Thursby *et al*., 1997; Phillips *et al*., 2001) indicate that the prior establishment of relevant effect levels is convenient for some ecotoxicological methods. The latter approach is more practical since the judgement of the appropriateness of analytical data is less laborious and is based on historical results.

In the present study, relevant effect levels were established for 11 ecotoxicological methods with different test organisms (*Daphnia similis*, *Ceriodaphnia dubia*, *Hyalella azteca*, *Hyalella meinerti*, *Danio rerio*, *Lytechinus variegatus* and *Mysidopsis juniae*), using Minimum Significant Difference (MSD) values, as suggested by Denton & Norberg-King (1996). Such relevant effect levels were also used to demonstrate their appropriateness in bioequivalence hypothesis testing, as suggested by Chapman *et al.* (1996).

MATERIALS AND METHODS

The calculation of the relevant effect levels for each test method was based on data from assays with single chemicals, industrial effluents and environmental samples performed in the Aquatic Ecotoxicology Laboratory – CETESB (São Paulo State, Brazil), as described in Table 1.

Initially, normality tests (Shapiro-Wilks and χ^2) were conducted for all toxicity test results. The homocedasticity of the data was then evaluated through Barttlet, Hartley and Levene tests (in case of effluents and chemicals) or F-test (in case of environmental samples – surface or interstitial waters and sediments). Thereafter, a multiple comparison between treatments and control means was made through Dunnett's test with or without Bonferroni adjustment for assays with several sample dilutions (effluents and chemicals), while t-test was applied for environmental samples. Analyses were performed with the Software TOXSTAT 3.5 (West Inc. & Gulley, 1995).

Table 1 – Organisms, number of tests and experimental conditions.

Test organisms (endpoint)	Number of tests	Number of dilutions*	Number of replicates	Methods
Daphnia similis (immobility)	41	2	5	ABNT (2004)
Daphnia similis (immobility)	101	6	4	ABNT (2004)
Mysidopsis juniae (survival)	118	6	3	ABNT (2005b)
Hyalella meinerti (survival)	87	2	4	ABNT (2007)
Hyalella azteca (survival)	36	2	4	ABNT (2007)
Danio rerio (larval survival)	43	6	4	Bertoletti (2000)
Danio rerio (embryolarval survival)	41	6	4	Bertoletti (2000)
Ceriodaphnia dubia (reproduction)	102	2	10	ABNT (2005a)
Ceriodaphnia dubia (reproduction)	42	6	10	ABNT (2005a)
Lytechinus variegatus (embryolarval development)	92	2	4	ABNT (2006)
Lytechinus variegatus (embryolarval development)	75	6	4	ABNT (2006)

* Including experimental control (2 = toxicity tests with ambient water or sediment; 6 = toxicity tests with effluents or single chemicals).

In both cases, generated MSD values were registered as a percentage of the control mean, which corresponded to the following calculation:

$$\%MSD = \frac{MSD}{\text{control mean}} \times 100$$

MSD percentage values were arranged in increasing order for characterization of their distribution and comparison with other data sets. The 10th, 25th, 50th, 75th and 90th percentiles were identified. The bioequivalence constants (B) were then calculated by deducting the value corresponding to the 75th percentile from 100, following the procedure adopted by Phillips *et al*. (2001). The bioequivalence hypotheses tests were performed with the Software TOXSTAT 3.5 (West Inc. & Gulley, 1995).

RESULTS AND DISCUSSION

The establishment of significant adverse effect levels in a population of aquatic organisms is difficult, mainly due to the lack of ecological studies designed with this objective. At the same time, this information has become valuable in the ecotoxicological assays used to provide information relevant for the protection of aquatic life.

Some researchers state that the choice of significant effect level is arbitrary and depends on biological, statistical, social and regulatory issues (Garrett, 1997; Shukla *et al*., 2000). Thursby *et al*. (1997) consider these effect levels as thresholds and mention that "the exact method of determining a threshold is not as important as having such a threshold".

Due to such difficulty, Denton & Norberg-King (1996) suggested the establishment of the practical significance level, in substitution to the biological significance level. The practical significance level is based on minimum significant difference (MSD) values, calculated through classical hypothesis testing applied to ecotoxicological assays results. Some researchers have already confirmed the convenience of using MSD limits as criteria for the acceptance of variance in such assays (Thursby *et al*., 1997; Wang *et al*., 2000; Phillips *et al*., 2001).

Considering these principles, the present study establishes the relevant effect levels for different ecotoxicological methods, with single treatment and multiple concentrations, using several aquatic organisms and endpoints (Figure 1). The established levels are also based on the technique suggested by Thursby *et al*. (1997) and Phillips *et al*. (2001), which considers the variance of the whole assay (control and sample), rather than preferential variance of experimental controls, as suggested by Wang *et al*. (2000). Although the estimates of the relevant effect levels are obtained by statistical methods, they are originary from historical data and allow the evaluation of reliable variability of the analytical methods.

In the present study, as suggested by Chapman *et al*. (1996) and demonstrated by Phillips *et al*. (2001), the MSD percentages differ according to the assay method (Figure 1). Therefore, considering 10th and 90th percentiles of multiple concentration acute tests (Figure 1A), the *Mysidopsis juniae* assay presented the highest variation (7 and 29%), followed by *Daphnia similis* (13 and 34%). The lowest variations were found in single concentration tests with *Hyalella azteca* (4 and 15%) and *Daphnia similis* (6 and 19%).

The highest variation of chronic tests (Figure 1B), in terms of 10th and 90th percentiles, was observed in embryolarval *Danio rerio* toxicity tests (5 and 29%) with multiple concentrations, followed by the same type of assay with *Ceriodaphnia dubia* (16 and 37%). Conversely, the lowest values were obtained in single concentration embryolarval development tests with sea-urchin *Lytechinus variegatus* (3 and 12%). As a general rule, single concentration toxicity tests showed less variance than multiple concentration assays.

These values are similar to those registered in a survey conducted by USEPA (2000) using the corresponding test protocols, in which results obtained by various laboratories with reference toxicant tests and multiple concentrations were combined (Table 2). The only exception is the *Daphnia* test, where the variability with the species utilized in this study (*Daphnia similis*) was higher than with *Daphnia magna* and *Daphnia pulex* (Table 2).

The value of 90th percentile of the MSD for the *Lytechinus variegatus* assay exhibited low variability (12%) relative to the value obtained (22%) with the equinoderm *Strongylocentrotus purpuratus* (Phillips *et al*., 2001), despite a lower number of replicates. These data demonstrate that, although similar in the experimental design, some variability of response can be achieved within a taxonomical group, as observed with the genera *Hyalella* (Figure 1A) and *Daphnia* (Figure 1A and Table 2).

The percentiles shown in Figure 1 and Table 2 are equivalent to the statistical power of the assays, indicating the frequency of results obtained after several repetitions. The definition of the statistical power allows the restriction of the occurrence of false results and, indirectly, reduces the assay variability by improving its execution in a given laboratory.

On the other hand, the choice of the statistical power, which has been arbitrarily established, can indirectly reflects the level of environmental protection intended. For instance, Thursby *et al*. (1997) used a power of 80%, while Phillips *et al*. (2001) selected 90%. We considered that 75% would be the most appropriated statistical power (Table 3) because it offers greater environmental protection as long as minor variability is allowed. Although the power of 75% increases the chance of non-compliance by laboratories, USEPA (2000) has demonstrated that such statistical power can be obtained by several laboratories. These arguments can be used by regulators to justify the choice of 75th MSD percentile as practical, and consequently of biological relevance. Therefore, when the ecotoxicological assay does not reach this power, it is implicit that the laboratory needs to improve its analytical performance and to repeat the assay. This improvement can be achieved by increasing the number of replicates, increasing the number of organisms per treatment, improving the culture conditions and training technicians.

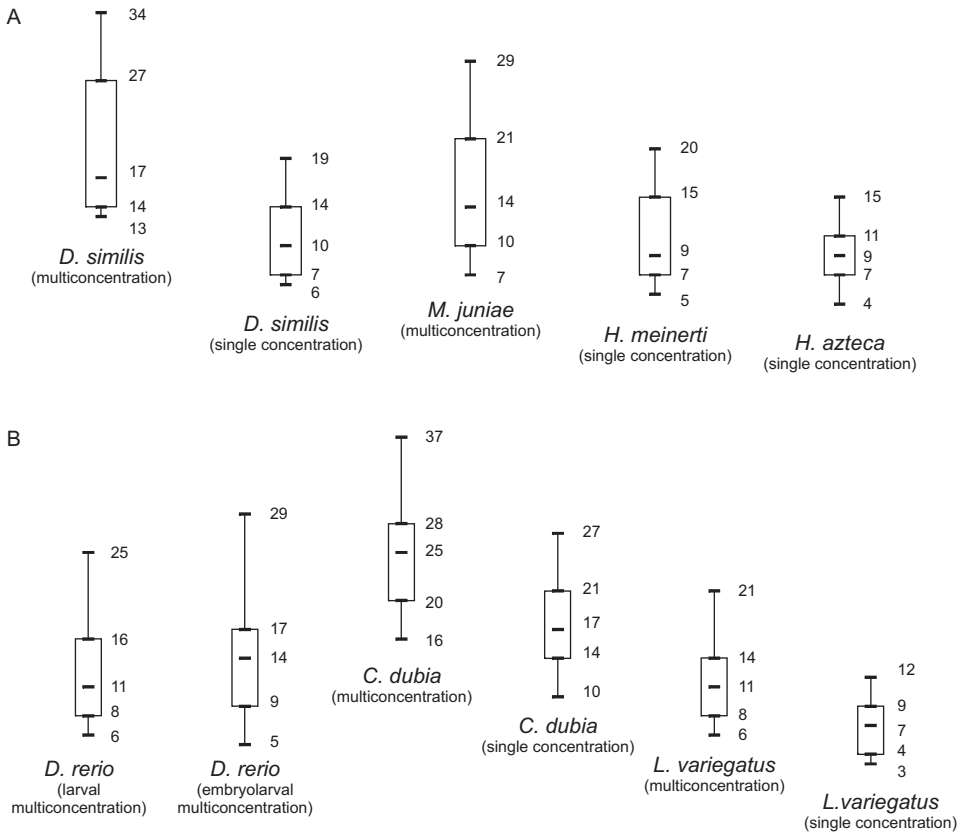

Figure 1 – MSD percentages distribution for different acute (A) and chronic (B) assay methods. Lower and upper limits of boxes correspond to 25th and 75th percentiles, respectively. Horizontal line inside each box represents the 50th percentile. Maximum and minimum values are 10th and 90th percentiles.

Table 2 – Variability of Minimum Significant Difference (MSD) at different statistical powers for various organisms and multiple concentrations test methods (USEPA, 2000).

Test organism/Endpoint	MSD (as %) at different percentiles				
	10th	**25th**	**50th**	**75th**	**90th**
Ceriodaphnia dubia (reproduction)	11	16	23	30	37
Pimephales promelas (survival)	9	14	20	25	35
Daphnia magna (survival)	5.3	8.4	13	19	23
Daphnia pulex (survival)	5.8	8.3	14	20	23
Americamysis bahia (survival)	5.1	8.9	15	23	26

Denton & Norberg-King (1996) also calculated the level of sensitivity for a series of standard procedures. Considering 150 data sets for *Ceriodaphnia dubia* reproduction in chronic toxicity tests with sodium chloride, these authors verified that the 75th MSD percentile corresponded to 7.5 neonates. Such number was correspondent to 32% of the average of neonates produced in the control (23.6), which is close to the value obtained in the present study (28%).

The values corresponding to the 75th percentile of MSD (Table 3) allow laboratories, as well as the environmental agencies, to verify the analytical quality of ecotoxicological tests results. Hence, when the classical hypothesis test ("t")

is used, the minimum significant differences should be smaller than the percentile described in the present study to avoid results with high variability. Therefore, if a result presents larger variability than the percentage described in Table 3, the assay should be repeated. In the same way, these 75th percentiles of MSD can be used in the statistical method of point estimates, just like the linear interpolation (Buratini & Bertoletti, 2006), which is an alternative statistical analysis method for chronic toxicity assays with some organisms.

The use of MSD limits has also been recommended by USEPA (2000) in the regulatory context, with the purpose of verifying the acceptable variability of hypothesis testing with experimental control and single treatments. The approach recommended by USEPA (Table 2) allows variation of the MSD within the previously established limits (among the 10th and 90th percentile). However, the judgement of the acceptable variation depends on the decision maker in a given environmental agency, as well as on the occasional analytical performance of the laboratory conducting the assay. Difficulties in such judgment were found by Phillips *et al*. (2001) when t-test results and MSD were used without a direct integration of data. Such facts become more critical any time that the analytical results are requested for regulatory purposes.

The bioequivalence hypothesis testing has shown to be suitable to avoid these occurrences, due to the implicit incorporation of a relevant effect level as discussed by several authors (Erickson & McDonald, 1995; Garret, 1997; Gully *et al*., 2000; Shukla *et al*., 2000). Therefore, the association of bioequivalence hypothesis testing and previous values of B (bioequivalence constants) is useful to avoid inconstancy in judgment, as well as for prior establishment of the acceptable variability of the analytical result. It is suitable to point out that values of B (bioequivalence constants) are the complementary percentual of the 75th percentile of MSD, expressed as proportions, and for this reason they should be used when statistical calculation requires this kind of values.

As an actual example, using the B = 0.79 (Table 3) in bioequivalence hypothesis testing with data from *Ceriodaphnia dubia* assays, a reduction of incidence of false positives can be observed in Table 4. It is verified that samples from first and second assays produced 21 and 26 neonates per female, respectively. These values are much higher than test acceptance criteria (15 young per female) and near the mean reproduction in several laboratories (23.2 offspring per female) reported by Moore *et al*. (2000). Consequently, such results could be considered as lacking biological difference. However, only bioequivalence tests identified them as non toxic, because the relevant effect levels for assays 1 and 2 were 20 and 24 neonates respectively, while classical hypothesis testing (t-test) generated false positives.

Table 3 – Relevant effect levels (75th MSD percentiles and bioequivalence constants) for different test methods.

Test organisms	Number of dilutions*	Number of replicates	75th MSD percentiles (%)	Bioequivalence constants (B)
Daphnia similis (immobility)	2	5	14	0.86
Daphnia similis (immobility)	6	4	27	0.73
Mysidopsis juniae (survival)	6	3	21	0.79
Hyalella meinerti (survival)	2	4	15	0.85
Hyalella azteca (survival)	2	4	11	0.89
Danio rerio (larval survival)	6	4	16	0.84
Danio rerio (embryolarval survival)	6	4	17	0.83
Ceriodaphnia dubia (reproduction)	2	10	21	0.79
Ceriodaphnia dubia (reproduction)	6	10	28	0.72
Lytechinus variegatus (embryolarval development)	2	4	9	0.91
Lytechinus variegatus (embryolarval development)	6	4	14	0.86

* Including experimental control (2 = toxicity tests with ambient water or sediment; 6 = toxicity tests with effluents or chemicals).

Table 4 – Results of classical ("t") and bioequivalence hypothesis testing, applied to the means of neonates registered in *Ceriodaphnia dubia* chronic assays with surface water samples.

Assay	Mean number of neonates		Relevant effect level[A] (79% of control neonates)	Conclusion of statistical analysis	
	Control	Sample		t-test	Bioequivalence test (B = 0.79)
1	26	21	20	Toxic	Non toxic
2	31	26	24	Toxic	Non toxic
3	17	11	13	Toxic	Non toxic
4	27	10	21	Toxic	Toxic
5	34	31	26	Non toxic	Non toxic

A = number of organisms below which there is a significant difference.

Assay number 3 showed significant toxic effect by using t-test, but such significance was not observed with the bioequivalence test, although the mean of neonates in the sample was below the relevant effect level. Such result suggest a failure of the bioequivalence test, but data analysis indicated that this was due to the high power of the t-test result in this particular experiment (MSD < 9.3%), which has a low probability of occurrence throughtout the time for this method (<10%, as showed in Figure 1B). Therefore, a false positive result was obtained in this experiment when using classical hypothesis testing.

The mean number of neonates relative to the sample from the fourth assay was much lower than the mean obtained in the control and was identified as significant by both statistical approaches, thus reflecting effects of biological relevance. Finally, the fifth assay, which clearly did not exhibit either significant statistical or biological difference, was considered non toxic using both approaches.

The relevant effect levels presented in this work are specific for the methods and test organisms used. Additional calculations are unnecessary whether the experimental conditions are similar to those described in the present paper or conditions recommended in ABNT's (Associação Brasileira de Normas Técnicas) standard procedures are used. However, the calculation of the bioequivalence tests through specific software (as used in this study) is recommended, in order to save time and laborious calculations.

As was demonstrated, the integration of relevant effect levels and bioequivalence hypothesis testing is appropriate for the calculation of significance of ecotoxicological assays results (Table 4), whether with single or multiple concentrations (data not presented). This feature allows the avoidance of the frequent occurrence of false positives observed when classical hypotheses tests are applied to such data. This statistical approach deems unnecessary the prior negotiation of laboratory performance between regulated and regulatory agencies (as proposed by Erickson & McDonald, 1995), and avoids inconstancy in judgment for the acceptance of variability of analytical results.

REFERENCES

ABNT (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS), 2004, *Ecotoxicologia Aquática – Toxicidade aguda – Método de ensaio com Daphnia spp (Crustacea,Cladocera)*. ABNT-NBR 12713, 21p, Errata 1.

ABNT (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS), 2005a, *Ecotoxicologia Aquática – Toxicidade crônica – Método de ensaio com Ceriodaphnia spp (Crustacea,Cladocera).* ABNT-NBR 13373, 15p.

ABNT (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS), 2005b, *Ecotoxicologia Aquática – Toxicidade aguda – Método de ensaio com misidáceos (Crustacea).* ABNT-NBR 15308, 17p.

ABNT (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS), 2006, *Ecotoxicologia Aquática – Toxicidade crônica de curta duração – Método de ensaio com ouriço-do-mar (Echinodermata: Echinoidea).* ABNT-NBR 15350, 17p.

ABNT (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS), 2007, *Ecotoxicologia Aquática – Toxicidade em sedimento – Método de ensaio com Hyalella spp (Amphipoda).* ABNT-NBR 15470, 20p.

BERTOLETTI, E., 2000, *Estimativa de efeitos tóxicos crônicos com Danio rerio (Pisces: Cyprinidae)*. Tese de Doutorado, Faculdade de Saúde Pública da USP, São Paulo, 120 p.

BURATINI, S. V. & BERTOLETTI, E., 2006, Análise estatística. In: P. A. Zagatto & E. Bertoletti (eds.), *Ecotoxicologia Aquática: Princípios e aplicações*. RiMa, São Carlos/SP, pp. 221-249.

CHAPMAN G. A., ANDERSON, B. S., BAILER, A. J., BAIRD, R. B., BERGER, R., BURTON, D. T., DENTON, D. L., GOODFELLOW, W. L., HEBER, M. A., MCDONALD, L. L., NORBERG-KING, T. J. & RUFFIER, P. J., 1996, Methods and Appropriate Endpoints. In: D. Grothe, K. L. Dickson & D. K. Reed-Judkins (eds.), *Whole Effluent Toxicity Testing: An evaluation of methods and prediction of receiving system impacts*. SETAC Press, Pensacola, Florida, pp. 51-82.

DENTON, D. L. & NORBERG-KING, T. J., 1996, Whole Effluent Toxicity Statistics: a Regulatory Perpective. In: D. Grothe, K. L. Dickson & D. K. Reed-Judkins (eds.). *Whole Effluent Toxicity Testing: An evaluation of methods and prediction of receiving system impacts*. SETAC Press, Pensacola, Florida, pp. 83-102.

ERICKSON, W. P. & MCDONALD, L. L., 1995, Tests for bioequivalence of control media and test media in studies of toxicity. *Environ. Toxicol. Chem., 14*(7): 1274-1256.

GARRET, K. A., 1997, Use of statistical tests of equivalence (bioequivalence tests) in plant pathology. *Phytopathology, 87*(4): 372-374.

GULLY, J. R., BAIRD, R. B., MARKLE, P. J. & BOTTOMLEY, J. P., 2000, Effect-based interpretation of toxicity test data using probability and comparison with alternative methods of analysis. *Environ. Toxicol. Chem*., *19*(1): 133-140.

MOORE, T. F., CANTON, S. P., GRIMES, M., 2000, Investigating the incidence of type I errors for chronic whole effluent toxicity testing using *Ceriodaphnia dubia. Environ. Toxicol. Chem*., *19*(1): 118-122.

PHILLIPS, B. M., HUNT, J. W., ANDERSON, B. S., PUCKETT, H. M., FAIREY, R., WILSON, C. J. & TJEERDEMA, R., 2001, Statistical significance of sediment toxicity test results: threshold values derived by the detectable significance approach. *Environ. Toxicol. Chem., 20*(2): 371-373.

SHUKLA, R., WANG, Q., FULK, F., DENG, C. & DENTON, D., 2000, Bioequivalence approach for whole effluent toxicity testing. *Environ. Toxicol. Chem*., *19*(1): 169-174.

THURSBY, G. B., HELTSHE, J. & SCOTT, K. J., 1997, Revised approach to toxicity test acceptability criteria using a statistical performance assessment. *Environ. Toxicol. Chem., 16*(6): 1322-1329.

U.S.EPA (United States Environmental Protection Agency), 2000, *Understanding and Accounting for Method Variability in Whole Effluent Toxicity applications Under the National Pollutant Discharge Elimination System Program.* EPA/833/R-00/003. U.S.EPA, Washington, DC, p.i.

WANG, Q., DENTON, D. L. & SHUKLA, R., 2000, Application and statistical properties of minimum significant difference-based criterion testing in a toxicity testing program. *Environ. Toxicol. Chem*., *19*(1):113-117.

WEST INC. & GULLEY, D., 1995, *TOXSTAT 3.5*. University of Wyoming, Wyoming, USA.