N00129.AR.001029
NSB NEW LONDON
5090.3a

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
REGION 1
1 CONGRESS STREET, SUITE 1100
BOSTON, MASSACHUSETTS 02114-2023

March 3, 2004

Mark Evans, Remedial Project Manager
U.S. Department of the Navy
Naval Facilities Engineering Command
Northern Division
10 Industrial Highway
Code 1823, Mail Stop 82
Lester, PA 19113-2090

Re: Thames River - Basis for Data Quality Objectives and Survey Design (Report 1) & Baseline Ecological Risk Assessment Sampling Design (Report 2)

Dear Mr. Evans:

EPA reviewed the Basis for Data Quality Objectives and Survey Design (Report 1) and the Baseline Ecological Risk Assessment Sampling Design (Report 2) for the Thames River investigation that is part of the Lower Subase at the Naval Submarine Base. This letter focuses on the statistical methods proposed in these reports and supplements EPA's letter dated February 27, 2004. Since, there is a considerable overlap in the contents of the two reports, the following general and specific comments are applicable to both reports. Only statistical issues about the proposed methods have been discussed in this review letter report.

While the sequence and mechanism of the proposed statistical methods to develop sampling survey designs and to assess ecological risks are generally suitable for the intended purpose of developing appropriate sampling designs for the baseline ecological risk assessment, several of the assumptions made (normality assumption, equality of variances assumption) to be able to use the proposed statistical tests (*e.g.*, Dunnett test, sample size determination formulae) may not be tenable by the data collected (*e.g.*, survival rate of amphipod *Leptocheirus plumulosus*, or the number of offsprings per surviving adult in 28-d lab experiments) from the three areas of concern (AOC) (and also from the reference area). In that case, the use of non-parametric tests (as mentioned in EPA 2001 document) such as the Wilcoxon Rank Sum Test is preferable (*see also* Vollandt and Horn, 2000). Also, some sample size equations are in error. Furthermore, the various hypotheses and decision rules as described in Step 5 (Develop Decision Rule(s)) of the various DQOs (Tables 1-3, Report 1) are not properly stated and tested.

In these two documents, the Navy proposes use of the Dunnett test, therefore it is desirable to include this procedure and its functionality in brief in at least one of these two reports. This will greatly benefit reviewers of the two reports. The Dunnett test is a reasonable test to compare a finite number of treatments (represented by the three lower NSB locations in the present study) with a single control (*e.g.*, the reference area) treatment provided the assumptions needed are

satisfied. Furthermore, there seems to be some confusion about the appropriate use of the Dunnett test for the present applications as can be seen from Step 5 (Develop a Decision Rule) of the DQOs given in Tables 1-3 of Report 1. A good description of the proposed test procedure (with examples) should clarify its use. All relevant details (*e.g.*, for sample size calculations) with references should be included in at least one of the two reports. Some specific comments on the proposed statistical procedures and underlying assumptions are discussed in detail below.

The two reports proposed the use of several statistical methods to determine appropriate sediment sample sizes (using equation 1, page 5, Report 2), number of replicates of species (*e.g.*, amphipods *Leptocheirus plumulosus)* lab bioassay tests, and the number of forage fish composite tissue samples *etc.* (using equation 2, page 5, Report 2). The data will be used to assess the ecological risks (via *Leptocheirus plumulosus* and birds consuming forage fish) from contaminants in the three AOC. Samples need to be collected from a reference area (control) and the three areas (k=3 treatments) of concern (zones 4, 7, and Pier 1). The three treatment (AOC) data on sediment concentrations, fish tissue concentrations, and species survival characteristics (survival, growth, reproduction in 28-d lab experiments) collected from the three AOC will be compared with the reference (previously agreed upon general background) area sediment concentrations, fish tissue samples, and species characteristics (observed in 28-d lab tests). The use of Dunnett test has been proposed for all such comparisons.

Even though the statistical methods as proposed in the two reports have been available in the statistical literature for quite some time (*e.g.*, since 1955 - Dunnett test- often used in biological applications), these methods [*e.g.*, Dunnett test and associated sample size equation 1 (Report 2)] represent relatively new methods in environmental applications such as those proposed herein. The actual appropriate use and benefit (in terms of better performance, higher power, *etc.*) of these statistical methods may not be entirely clear to a typical reviewer.

The sample size equation (2) as given in Report 2 (page 5) is not correct. Since the means of two populations are being compared, there should be a factor 2 in the numerator of equation (2), page 5 (*e.g.*, Zar (1984)). This equation (2) has been used to derive number of replicates for the species bioassay tests, and also to determine number of composite fish tissue samples. These sample sizes should be recomputed using the correct equation. Obviously, the actual sample sizes will be twice the sample sizes as derived in the two Navy reports.

On page 6 of Report 2, it is stated that α, β, and power as described on pages 5 and 6 are applicable to the two sample t-tests. However, use of Dunnett's test for treatment versus control comparisons is proposed in the present study. Therefore, instead of describing α and β for t-test, the report should include these definitions for Dunnett's test. On page 6, it is stated that, "definitions for α and β are modified to include interpretations based on rejecting at least one of the paired comparisons." Please include these modified definitions and actual values used in the present report (Report 2) and clarify how the α, β, and power differ for the proposed Dunnett's test from the traditional two sample t-test.

For all of the hypotheses tested (as described in Tables 1-3 of Report 1, and also described in Report 2), the same sample size is used for all of the three treatment groups and also for the control group. However, k (3) treatment means are compared with the mean of a single control group. Dunnett (1955) recommends a higher sample size for the control group. Dunnett actually showed that the optimal sample size (to achieve specified power) for the control group is about $n0 \approx n\sqrt{k}$, where n is the common sample size for the k treatment samples. For example, if samples of size 5 each are collected from the three treatment groups (using equation (1), page 5, Report 2), then the optimum value for the sample size from the reference area will be about 9 ~ 5x1.732. This option should be investigated.

The reference for equation (1) given on page 5 should be included in Report 2.

The report states that all hypotheses will be tested using the Dunnett test, but only the sample sizes to determine the sediment chemistry of the COPC will be computed using equation (1), the sample size equation derived for the Dunnett test. Equation (2) is not correct and it is therefore not clear why it is used to determine sample sizes to assess sediment toxicity using 28-d bioassay test replicates and forage fish tissue samples. Equation (2) is typically used to determine sample sizes for two sample t-tests (*e.g.*, Zar, 1984). Please explain.

The Dunnett test as used in the two reports (Dunnett, 1955) is meant to compare the *means* of some k (finite number) treatment populations with the *mean* of the same control treatment (reference area). This test will change when one tries to compare the proportions (*e.g.*, the proportion of surviving amphipod *Leptocheirus plumulosus* at the end of 28-day controlled laboratory experiment as defined on page 13 of EPA 2001) of k treatments with the proportion of the control group. The normality assumption will not be satisfied by the proportions, especially when one is dealing with small sample sizes (*e.g.*, <10) as is the case here. It is preferable to use a suitable non-parametric test such as the Wilcoxon Rank Sum Test when dealing with proportions. The associated sample size equation and the sample size will also change. The sample size derived and proposed for comparing the *proportions* of surviving amphipod *Leptocheirus plumulosus* from the 3 AOC may not be appropriate (as it is based upon a two sample t-test to compare the means), and it may not meet the DQOs.

Some researchers (*e.g.*, Vollandt and Horn) have worked on the problem of determining the sample sizes when comparing proportions of k treatments with the proportion of a single control treatment. It is desirable to investigate the applicability of the sample size formula for comparing the proportions as derived by Vollandt and Horn. Also some alternative non-parametric methods are provided in the sample size determination manual prepared by Vollandt and Horn (2000). Some of those non-parametric methods may be applicable in the present applications, for example when one is comparing proportions of surviving adult amphipod *Leptocheirus plumulosus.*

Similarly, the variable number of offsprings of a surviving adult amphipod *Leptocheirus plumulosus* at the end of the 28-d lab bioassay test will not follow a normal distribution violating

the normality assumption needed to use the Dunnett test. Accordingly, an appropriate test needs to be developed to test the reproduction hypotheses for the three AOC. The associated sample size equation and the sample size will also change.

Three different AOC are being compared with a reference area. It is likely that the levels of contamination at the three AOC are different, consequently the variabilities of the three AOC and the reference area may be significantly different. In this set up (using AOC as the treatments), it is hard to assume or justify the equality of variances of all treatment populations (areas of concern) under consideration including the control reference area (or lab controlled sediment results *etc.*) population. This equality of variances is one of the fundamental assumption that is needed to perform Dunnett test (because MSE derived for One-way Analysis of Variance is used in the Dunnett test) and also to determine appropriate sample sizes (equation (1), page 5, Report 2) for Dunnett test. It should be noted that the equality of variances is also needed to be able to use the sample size equation (2) of Report 2. The violation of this assumption of equality of variances could lead to incorrect results (sample size), inference, and conclusions.

In Dunnett test, k null hypotheses (*e.g.*, treatment mean = control mean, for each of the k treatments), and k alternative hypothesis (*e.g.*, treatment mean > control mean), one for each treatment are defined. Thus, k pair-wise comparisons are made comparing the k (3) treatment means with the control mean. There is one critical value associated with each such pair-wise comparison. These critical values are available for 0.05 and 0.01 levels of significance (Dunnett, 1955). The observed values of the Dunnett test are compared with the associated respective critical values (*e.g.*, Zar). If the computed value exceeds tabulated critical value (*e.g.*, at 0.05 level of significance), the corresponding null hypothesis is rejected.

There is no one overall composite critical value for the Dunnett's test as suggested in the two reports (*e.g.*, as described in Step 5 of Tables 1 -3, Report 1). If there are such critical values available for the Dunnett test, those tables and references should be included in at least one of the reports.

On page 4, it is stated that, " For any given chemical, is if the p-value from Dunnett's test is less than or equal to 0.1, the null hypothesis will be rejected and concluded that at least one of the mean concentrations from Thames River Lower Subbase sites is greater than the mean concentration at the Reference Area. Subsequently, for any given chemical, each pair-wise comparison is then evaluated to determine which mean chemical concentrations are greater than the Reference Area." The Dunnett test actually performs these pair-wise comparisons using a common estimate of the variability (the MSE from One-way Analysis of Variance), which of course requires the equality of variances of all treatments including the control treatment group.

The proposed Dunnett's test also requires the normality of the data from all treatment populations (Zones 4, 7, and Pier 1) and the control population (lab control conditions or reference area). Obviously this assumption can not be satisfied by the survival rate (defined as the proportion of surviving amphipod *Leptocheirus plumulosus)* and the reproduction (defined as

the number of offsprings of the surviving adult). Since we are dealing with small samples (5-6 lab replicates), it may not be appropriate to approximate proportions by a normal distribution. As mentioned before, some non-parametric test will be better suited to compare the survival rate and reproduction characteristic of the three AOC with that of the lab controlled and reference area results.

Some of the researchers on the current topic of pair-wise comparisons performed power comparisons using the Dunnett test and also using the pair-wise two sample t-test. They concluded that the pair-wise two sample t-test had higher power than the Dunnett's test in differentiating between the means of two normal populations that are being compared. This is an interesting observation that may warrant further investigation. However, for the present applications, it may be worth performing (can be programmed) both tests: the Dunnett test and the two sample t-test (when the assumptions are met) to perform pair-wise comparisons while comparing k treatments with the same control.

I look forward to working with you and the Connecticut Department of Environmental Protection to protect the Thames River. Please contact me at (617) 918-1385 to arrange a conference call to discuss these comments.

Sincerely,

Kymberlee Keckler, Remedial Project Manager
Federal Facilities Superfund Section

cc: Mark Lewis, CTDEP, Hartford, CT
Melissa Griffin, NSBNL, Groton, CT
Bart Hoskins, USEPA, Boston, MA
Jennifer Stump, Gannett Fleming, Harrisburg, PA

REFERENCES:

EPA 2000. Guidance for Data Quality Assessment. Practical Methods for Data Analysis. EPA QA/G-9. EPA/600/R-96/084.

EPA 2001. Methods for Assessing the Chronic Toxicity of Marine and Estuarine Sediment-associated Contaminants with amphipod *Leptocheirus plumulosus* , EPA 600-R-01-020. USEPA - ERLIN, Office of Research and Development, Washington D.C.

Dunnett, C.W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of American Statistical Association*, 50, 1096-1121.

Vollandt, R., and Horn, M. Sample Size Determination for Multiple Many-One and Pair-Wise Comparisons of Proportions Using the Arcsin Transformation.

Vollandt, R., and Horn, M (2000). A Manual for the Determination of Sample Sizes for Multiple Comparisons - Formulas and Tables.

Zar, J. H.(1984). Biostatistical Analysis, Second Edition. Prentice Hall.