N00129.AR.001137
NSB NEW LONDON
5090.3a

Response to July 19, 2005 Letter
U.S. Environmental Protection Agency, Region I

The comments presented by EPA Region I indicate that there is some remaining confusion relating to the New London Lower Subase Workplan and the ecological risk data interpretation process. The following responses are intended to further clarify these points and to set the stage for continued discussions and resolution on the next steps necessary for risk decision making at the Lower Subase sites.

Comments from U.S. EPA Letter (July 19, 2005):

1. **EPA extensively reviewed the data from the Thames River toxicity sampling in consultation with the developers of the toxicity test method. This letter addresses the specific question of how to use the existing data, and whether there is a need for further testing. In addition, EPA raises new concerns about how any additional test data will be evaluated with respect to laboratory controls and reference areas.**

   As you know, the test was impaired by the failure to meet laboratory control performance standards, which raises fundamental questions about the initial health and condition of the test organisms. Also of concern is the extreme degree of variability within replicates of control, reference, and site samples. This variability may reflect a problem with initial test organism condition, or could indicate a problem with the test protocols and procedures related to randomization, homogenization of samples, or test conditions between replicates. The Navy did not state whether a concurrent reference toxicant test was performed, so EPA therefore assumes that there was no such concurrent test. The reference test could shed light on the condition of the test organisms.

   Response: As the Navy reported to EPA the bioassay data for the laboratory control did not meet two of the EPA test protocol performance criteria. First, 80% or greater average survival of the laboratory control treatment was not met and second, that there must be no single replicate with survival of less than 60%. The bioassay for the laboratory control exhibited 76% average survival, which was driven by 2 of the 9 replicates that exhibited less than 60% survival. All other aspects of the test were met, including the reference toxicant test. The reference toxicant data reflect that indeed the organisms used in the test were responsive to the test toxicant and further indicate that the test, not withstanding the failures recognized above, was functional.

   With respect to the test variability, the EPA *Leptocheirus* Bioassay Protocol (*Method for Assessing the Chronic Toxicity of Marine and Estuarine Sediment-associated Contaminants with the Amphipod Leptocheirus plumulosus*, EPA 600/R-01/020, March 2001) recognizes that large variability does occur for the growth and reproductive test endpoints. Previous test results were evaluated to assess the anticipated variability and develop a sample design that considered this variability. For this reason, the New London Work Plan (Battelle and Neptune and Co, 2004) study design specified that 9 replicates were necessary to address this variability. Significant bioassay test variability was anticipated and accommodated in the survey design and Data Quality Objective (DQO) process

Another source of the variability is the heterogeneity of the COPCs with respect to temporal ands spatial releases as well as the physiochemical properties of many of the COPCs. As illustrated in the conceptual site model COPC, releases from the New London Lower Subase have varied since the base opened in 1868. The nature of these historical releases include residual coal (PAHs) from the old coaling activities in the late 1800's to residual heavy fuel oil (hydrocarbons and PAHs) leakage, DDT pesticide applications, and more recently metals often associated with ships and their maintenance (As, Ba, Cr, Cd, Cu, Ni, Se, Pb and Zn) and the common use of PCBs. These COPCs are typically only slightly water soluble, and often in a particulate and globular in physical form. The physical forms of most of these COPCs do not homogenization completely without changing the physical form (grinding or emulsifying) of the substrate. The sediment material is well mixed as part of the sample handling protocol, however, by their nature these compounds may demonstrate the replicate to replicate variability observed. This is further supports the need for a larger number of replicates, so that these site sample conditions are adequately represented to support appropriate site decision making.

2. **EPA is aware that the Navy has studied this test to identify the source of this problem, and that no apparent reason for these problems has been identified. Such problems are not un-precedented with the *Leptocheirus* test, but it has been the experience of EPA that a repeat test can resolve these problems. It is possible that for some unknown reason the test organisms were not in suitable condition for testing. I therefore recommend that an additional test be run using all but a few of the same sample locations. In an attempt to use as much of the existing data as possible, EPA has studied this data set and proposes several samples that can reasonably be considered to be non-toxic on the basis of this data set, and on the basis of comparison to reference locations from other recent tests. One sample is proposed for consideration as a toxic sample on the basis of this test.**

   The criteria for omitting certain stations from further testing as non-toxic are as follows:

   - **High survival, and low variability in survival. The quantitative criterion being that no more than one replicate out of ten had less than 50% survival, and most replicates had 75% to 100% survival.**
   - **Relatively high mean amphipod weight. The quantitative criterion being that most replicates have mean weights above 1.5 mg.**
   - **High reproduction numbers and low variability in reproduction. The quantitative criterion being most replicates having 20 or more neonates.**

   On the basis of these criteria, the following stations are suggested to be considered non-toxic, and could be omitted from further testing: Zone 4, Station Z4-S2 and Pier 1, Station P5.

   Similarly, the data were examined to determine whether any sediments could reasonably be considered to be toxic on the basis of this test. This is a more difficult determination because many samples had an almost even split between replicates with very low survival,

and replicates that would meet the non-toxic criteria. Since the reason for the variability is not known, it is unclear what an "average" condition really should be. For example, if the variability was caused by incomplete homogenization, there is a possibility that the average condition would be a "low magnitude unacceptable risk," because more homogenization would mix the poor-performing sediment with the high performing sediment. On this basis, only one sample, Station Zone 4, Z4-S6, had very consistent low survival, low mean organism weight (all below 1 mg), and low reproduction (except for a single replicate). I suggest that this sample be considered toxic and omitted from further testing.

Response: The Navy's perspective regarding the use of this bioassay data set is that these data from this bioassay reflect similar deficiencies in test survival to data used by EPA Region I in another ecological risk assessment (Tetra Tech, 2005[1]). It is the Navy's desire to allow for a comparable use of the bioassay data at New London. The Navy appreciates the compromise offered by EPA; however, the recommended interpretation from EPA would suggest that some of the tests results are valid while the Navy recommends that the entire data set be used to assess toxicity. Furthermore, it is unclear what the scientific basis is for the selection criteria offered by EPA. For example, how is "low variability" assessed especially given the expectation of high variability reported by EPA in their testing protocol? Similarly, what is the basis for a mean weight of 1.5 mg or reproduction rates greater than 20 neonates?

In a risk assessment, it is the average COPC concentration result and/or bioassay replicate finding within an appropriate spatial habitat for the assessment endpoint(s) of concern that is of interest, not the singular value reported for a sediment analytical sample or a bioassay replicate. The important purpose of replicates, in all scientific investigations, is to gain insight in the variability of the decision sensitive components of the decision. The Navy believes and almost all ecological data supports that environmental samples by their vary nature are variable and this does not in itself dismiss the value of the test findings. This recognition of variability is verified by the US EPA *Leptocheirus* Bioassay Protocol (Table 13.3, page 88).

3. **It should be understood that if these determinations are accepted, the designations as toxic or non-toxic would be taken, without statistical analysis, into the findings of the ecological risk assessment. It is not useful to perform additional statistical analysis of this data set owing to the extreme variability. This leads the discussion to the additional concern regarding interpretation of the re-test data.**

Response: The Navy respectfully disagrees with these determinations for the reasons stated in the response to Comment 2. The basis for the quantitative criteria proposed by EPA is unclear, and the Navy believes that the entire data set can be used for toxicity evaluation, with the exception that no comparisons to laboratory control will be conducted.

1 TetraTech NUS, Industri-Plex Superfund SiteOperable Unit 2, MSGRP Remedial Investigation Report Draft Final, Woburn, Massachusetts, RAC Region I, March 2005.

4. **The Sampling and Analysis Plan (SAP) contains a disconnect in logic that was not noted in previous reviews. The SAP states that a determination of "low magnitude unacceptable risk" or "high magnitude unacceptable risk" would only be made for samples with significantly lower survival than both the laboratory control and the Thames River Reference Area survival, *and* survival less than or equal to 70% of mean Thames River Reference Area survival.**

   Response: This statement reflects a misunderstanding of the SAP. Section 3.2 of the Thames River Validation Study Work Plan uses 70% of the mean Reference Area survival as a cut-point for distinguishing between low and high magnitude **unacceptable** risk, not as a criteria for distinguishing between acceptable and unacceptable risk. As stated in the Work Plan, "If mean *L. plumulosus* survival at any Subase station is statistically different from mean laboratory control survival and mean Thames River Reference Area survival as defined in the DQOs, but is greater than 70% of mean Thames River Reference Area survival, then that station is classified as **low magnitude <u>unacceptable</u> risk.**" (page 28, 3rd bullet). Likewise, the fourth bullet states that any Subase station that "...is statistically different from mean laboratory control survival and mean Thames River Reference Area survival as defined in the DQOs, and is less than 70% of mean Thames River Reference Area survival, then that station is classified as **high magnitude <u>unacceptable</u> risk.**" In other words, if the Subase survival is statistically significantly different from Reference Area survival, this is classified as "an unacceptable finding". The 70% of Reference Area criteria is simply a cut-point for defining **magnitude** of unacceptability of the Reference Area results, and was in response to EPA's concern that low survival in a Reference Area would bias the toxicity test interpretation.

5. **There are a number of problems associated with this approach. The most egregious problem concerns the use of the "70% of Reference" approach. It is very possible that the Thames River Reference Area could, in a re-test, have a mean survival of 70%. If this were the case, a site station would only be considered "unacceptable" if it had survival of 50% or less. This is tantamount to using an LC50 as a reference toxicity value (RTV) and is completely inconsistent with EPA's approach to ecological risk assessment.**

   Response: Again, this represents a misinterpretation of what is written in the Work Plan. In any situation, the site station would be considered unacceptable as long as it is statistically significantly different than the Reference Area finding (Section 3.2 of the Work Plan, third and fourth bullets). In the example the reviewer has stated above, the site station would be classified as a **<u>high magnitude unacceptable risk</u>** finding if it had survival less than 50%. Survival greater than 50% would still be considered **unacceptable**, as long as the finding was significantly different than the Reference Area finding. However, it would be classed as a **<u>low magnitude unacceptable risk</u>** finding. According to Section 3.2 of the Work Plan, the only way to arrive at an "acceptable" risk finding is to **not** be significantly different than the Reference Area finding (Section 3.2, first and second bullets).

6. **The 70% criterion appears to be an over-extension of the statement in the SAP that in large-scale tests, over various seasons and over large areas, *Leptocheirus* has been found to have a**

survival of 68% or more. This statement from the SAP has not be traced to its source as part of this review, but it appears to be a statement supporting the broad applicability of the test over many types of conditions, not an endorsement of 68% reference control as a desirable outcome. Regardless of the source and purpose of this statement, it does not provide any justification for setting an acceptable risk level at 50% survival.

Response: It appears that the reviewer is confusing two different 70% criteria that were used in the development of the decision rules. The first paragraph of Section 3.2 of the SAP provides a basis for determining the **acceptability of the Reference Area** for use in the risk assessment. That paragraph states that as long as mean Reference Area survival is 70% or greater, **then** the Reference Area is suitable for conducting statistical comparisons to site data. If mean Reference Area survival is less than 70%, it perhaps suggests that an inappropriate Reference Area was chosen; therefore Reference Area comparisons would not be used in making risk determinations at the Lower Subase. This is intended to ensure that Lower Subase risk estimates are conservative by ensuring that Reference Area findings perform to an acceptable standard, and this standard is scientifically based upon the 68% survival finding over large scale tests mentioned in the comment.

The second 70% criteria is the **unacceptable risk magnitude criteria** discussed in the previous comment. This criterion is not derived in any way from the 68% survival finding from the large-scale tests mentioned in this comment. The selection of 70% **of the Reference Area** finding as a basis to distinguish low vs high magnitude risk was based upon professional judgment and the reasoning put forth in the example provided by the reviewer – namely that in a worst case scenario (Reference Area Survival equal to exactly 70%), seventy percent of the Reference Area survival would equal approximately 50%, and we wished to ensure that any survival below 50% would be classified as **high magnitude** risk. These classifications attempt to recognize that all risk is not equal, but is a matter of degrees, and that a mean survival of 10% is likely to have a greater impact on benthic populations than a mean survival of 60%.

7. **I suggest that the laboratory control be used as little as possible in test interpretation. The laboratory control is used to evaluate the test organism condition, test water, food, temperatures, techniques, and other variables that can affect the test outcome and confound any findings pertaining to chemical contamination. The data analysis should instead focus on the Thames River Reference Area only. To account for variability and overall degradation within the lower Thames River, no Thames River Reference Area samples are proposed for omission from the re-test. As a side note in support of re-running the test, the authors of the manual for this test indicated that 80% survival in the laboratory control was, in itself, a compromise position. The developers of the test felt that 90% would be a reasonable criterion for test acceptability based on their extensive testing, but they made a concession to use 80% in order to minimize the need to have repeat tests in broad application over diverse areas.**

The approach presented in the SAP, on closer inspection, represents significant deviations from accepted and established protocols presented in the EPA manual for this test. It also

completely fails to adequately address the potential for growth or reproduction effects. The following simplified approach is suggested for evaluating re-test data:

- **Establish that the test conditions and organism health was acceptable by evaluating the laboratory control with respect to test acceptability criteria.**
- **Provided the test is acceptable, follow the statistical methods provided in the EPA toxicity test manual to determine whether any site station means show a statistically significant difference in survival, reproduction, or growth from the Thames River Reference Area stations. A more detailed discussion of statistical approaches for this type of test is found in *Methods for Measuring the Toxicity and Bioaccumulation of Sediment-associated Contaminants with Freshwater Invertebrates, Second Edition*, U.S. EPA, Office of Water, March 2000 (EPA/600/R-99/064).**
- **Any station that shows a significant difference in survival, growth, or reproduction will be considered to pose some level of risk, and these findings will be carried into the overall ecological risk assessment for the site.**

Response: The Navy respectfully disagrees with the reviewer's statement that the approach presented in the SAP "represents significant deviations from accepted and established protocols presented in the EPA manual for this test". The only deviation from the published protocol is the use of 9 replicates per sampling location in lieu of the 5 replicates recommended in the protocol. This increase in the number of replicates was based on the Navy's review of the anticipated test variability and the desire to improve the statically power by better accounting for within treatment variability. The statistical protocols outlined in the SAP (Table 3-2, Step 5) for evaluating the bioassay data are consistent with the methods and references provided in Section 12 of the test protocol. The test protocol states "a control sediment is used to judge the acceptability of the test" (Section 12.2.2.1). It does not, however, state that the control sediment cannot be used to assist in the determination of magnitude of response of the test sediments, as this reviewer seems to be implying. Indeed, Section 12.2.3 of the protocol states "The purpose of the 28-d *L. plumulosus* chronic toxicity test is to determine whether the biological response to a treatment sample differs from the response to a **control (emphasis added)** or reference sample." For the Thames River SAP, the Navy proposed comparing to both control and reference samples to assist in identification of the **magnitude of risk** observed, and to provide an alternate point for comparison if the Reference Area performance was found to be unacceptably impacted. These multiple comparisons do not deviate from the published test protocol, which does not deal with determining **magnitude of response**, or **suitability of chosen Reference Area**.

The Navy also finds the reviewer's comment that the SAP "completely fails to adequately address the potential for growth or reproduction effects" to be inaccurate. Table 3-2, Step 5 on pages 23 and 24 of the SAP clearly lists the hypothesis tests for both the growth and reproduction endpoints, as well as the statistical protocols to be used. In addition, specific rules for the interpretation of the growth and reproduction data are presented on page 29 of the SAP, and these decision rules **clearly state** the conditions that will lead to findings of **no unacceptable risk, low magnitude unacceptable risk, and high magnitude unacceptable risk** for each of these endpoints.

The reviewer's "simplified approach" for evaluation of the re-test data is basically what the New London Work Plan already proposes, except that the Work Plan includes a criterion for evaluating the acceptability of the Reference Area for making such comparisons, and a criterion for defining magnitude of risk when unacceptable risk is determined to exist.

The Navy respects the perspectives provided by the reviewer and hopes that the responses to the reviewer's comments are seen as clarifying the points contained in the SAP while recognizing that deficiencies in the bioassay test do not totally restrict the utility of these data and their use in risk interpretation. The opportunity for a compromise proposed in the reviewer's comments is entirely consistent with the Navy's desires. Your review of this response for comments will hopefully provide a further opportunity for dialogue and an agreement that fully respects the perspectives and concerns of all parties to the risk decision. The Navy looks forward to a continuation of this dialogue.