

Freshwater Sediment Standards Ecology Responses to Science Panel Questions

Questions from the Science Panel related to bioassay-specific issues

1. Bioassays in general

a. Question/Issue: Is the growth test an appropriate surrogate for reproduction?

Response: Although there is not much available information on this topic, there are published peer-reviewed papers which indicate that growth and reproduction sensitivity is similar for *Hyalella*, which has a fully aquatic life cycle. The lack of greater sensitivity in the reproductive endpoint is likely in large part due to the variability observed in those endpoints. When confounded by a terrestrial adult life phase (chironomids, mayflies), the variability issues will likely become even more pronounced.

Citation: U. Borgmann, R. Néron and W. P. Norwood (2001). Quantification of bioavailable nickel in sediments and toxic thresholds to *Hyalella azteca*. Environmental Pollution 111(2)-189-198.

Abstract: Bioaccumulation and chronic toxicity of nickel (Ni) to *Hyalella azteca* in Ni-spiked sediments was strongly affected by the source of sediment used. The total range in LC50s on a sediment concentration basis ranged over 20 fold. Differences in Ni toxicity generally matched differences in Ni bioaccumulation, and toxicity expressed on a body concentration basis varied less than three fold. Body concentrations, therefore, provide a much more reliable prediction of Ni toxicity in sediments than do concentrations in the sediment. Ni in overlying water was also a reliable predictor of Ni toxicity, but only in tests conducted in Imhoff settling cones with large (67:1) water to sediment ratios. Overlying water LC50s for tests in beakers varied 18 fold. Sediment and body concentrations of Ni tolerated by *Hyalella* were slightly higher in cones than in beakers. Reproduction was not affected significantly by Ni at concentrations below the LC50 and 10-week EC50s for survival and biomass production (including survival, growth and reproduction) were only marginally lower than 4-week EC50s (survival and growth only).

Citation: Ingersoll, C. G., Brunson, E. L., Dwyer, F. J., Hardesty, D. K. and Kemble, N. E. (1998), Use of sublethal endpoints in sediment toxicity tests with the amphipod *Hyalella azteca*. Environmental Toxicology and Chemistry, 17: 1508–1523.

Abstract: Short-term sediment toxicity tests that only measure effects on survival can be used to identify high levels of contamination but may not be able to identify marginally contaminated sediments. The objective of the present study was to develop a method for determining the potential sublethal effects of contaminants associated with sediment on the amphipod *Hyalella azteca* (e.g., reproduction). Exposures to sediment were started with 7- to 8-d-old amphipods. On day 28, amphipods were isolated from the sediment and placed in water-only chambers where reproduction was measured on days 35 and 42. Typically, amphipods were first in amplexus at about days 21 to 28 with release of the first brood between days 28 to 42. Endpoints measured included survival (days 28, 35, and 42), growth (as length and weight on day 28 and 42), and reproduction (number of young/female produced from days 28 to 42). This method

was used to evaluate formulated sediment and field-collected sediments with low to moderate concentrations of contaminants. Survival of amphipods in these sediments was typically >85% after the 28-d sediment exposures and the 14-day holding period in water to measure reproduction. **Reproduction was more variable than growth; hence, more replicates might be needed to establish statistical differences among treatments.** Previous studies have demonstrated that growth of *H. azteca* in sediment tests often provides unique information that can be used to discriminate toxic effects of exposure to contaminants. Either length or weight can be measured in sediment tests with *H. azteca*. However, additional statistical options are available if length is measured on individual amphipods, such as nested analysis of variance that can account for variance in length within replicates. Ongoing water-only studies testing select contaminants will provide additional data on the relative sensitivity and variability of sublethal endpoints in toxicity tests with *H. azteca*.

2. Bioassay species selection

- a. Question/Issue: What is the definition of the health of the benthic community, and are the chosen bioassays/species a proven indicator of benthic community health?

Response: A major intent is to protect functions provided by a community, especially the prey base for upper trophic levels, which both bioassay species are important contributors to. Additionally, the two species represent two very different feeding guilds and life histories common in benthic invertebrates. The *Chironomus* test is performed on the post-hatch larval stage of this aquatic insect which burrows in- and feeds largely on organic deposits in the sediments. The *Hyaella* test uses the earliest stages of this shredder/grazer which feeds at the sediment surface and seeks protection by shallow burial in the sediments. The national peer reviewers were asked several questions regarding the appropriateness of the selected bioassays; their responses are attached (Peer reviewer responses to bioassay issues).

The bioassays Ecology selected are included in the list of assays used in the Great Lakes region. A publication by Burton et al. (1996) discussed several assays, both benthic and pelagic, that should be used to assess an area of concern. Similar to the State's Sediment Management Standards, Ecology has made the policy decision to use organisms intimately associated with sediments for the purposes of evaluating sediments. For benthic species, the Great Lakes region adds the mayfly larvae (*Hexagenia*) and a freshwater amphipod (*Diporeia*) to the standard *Hyaella* and *Chironomus* assays conducted in Washington. Neither *Hexagenia* nor *Diporeia* are commonly used outside of the Great Lakes program, in large part since they depend on wild-caught organisms; *Hexagenia* is highly seasonal in availability, and *Diporeia* has recently experienced population crashes.

- b. Question/Issue: How sensitive are these bioassays to various chemicals, particularly the chemicals typically found at freshwater sites?

Response: Sensitivity varies greatly between chemicals; a species may be quite sensitive to one group of compounds but highly insensitive to another. Given this, there are data that indicate the species Ecology selected are not consistently insensitive. Roman et al. (2007) found sublethal chronic endpoints for *Chironomus* were more sensitive than four

other benthic invertebrates, with *Hyalella* being the least sensitive. In the same study both *Hyalella* and *Chironomus* were in the middle of the group when lethal endpoints were considered. Other screening values (Threshold Effects Concentrations/Probable Effects Concentrations) included other endpoints in their calculations, including benthic community analysis. However, no such data is available for our region.

Citation: Roman YE, De Schampelaere KA, Nguyen LT, Janssen CR. (2007). Chronic toxicity of copper to five benthic invertebrates in laboratory-formulated sediment: sensitivity comparison and preliminary risk assessment. *Sci Total Environ.* 2007 Nov 15; 387(1-3):128-40.

Abstract: Five benthic organisms commonly used for sediment toxicity testing were chronically (28 to 35 days) exposed to copper in standard laboratory-formulated sediment (following Organization for Economic Cooperation and Development guidelines) and lethal and sub-lethal toxicities were evaluated. Sub-lethal endpoints considered were reproduction and biomass production for *Lumbriculus variegatus*, growth and reproduction for *Tubifex tubifex*, growth and emergence for *Chironomus riparius*, and growth for *Gammarus pulex* and *Hyalella azteca*. Expressed on whole-sediment basis the observed lethal sensitivity ranking (from most to least sensitive) was: *G. pulex*>*L. variegatus*>*H. azteca*=*C. riparius*=*T. tubifex*, with median chronic lethal concentrations (LC50) between 151 and 327 mg/kg dry wt. The sub-lethal sensitivity ranking (from most to least sensitive, with the most sensitive endpoint between parentheses): *C. riparius* (emergence)>*T. tubifex* (reproduction)=*L. variegatus* (reproduction)>*G. pulex* (growth)>*H. azteca* (growth), with median effective concentrations (EC50) between 59.2 and 194 mg/kg dry wt. No observed effect concentrations (NOEC) or 10% effective concentrations (EC10) for the five benthic invertebrates were used to perform a preliminary risk assessment for copper in freshwater sediment by means of (a) the "assessment factor approach" or (b) the statistical extrapolation approach (species sensitivity distribution). Depending on the data (NOEC or EC10) and the methodology used, we calculated a Predicted No Effect Concentration (PNEC) for sediment between 3.3 and 47.1 mg Cu/dry wt. This range is similar to the range of natural (geochemical) background concentrations of copper in sediments in Europe, i.e. 90% of sediments have a concentration between 5 and 49 mg Cu/kg dry wt. A detailed analysis of the outcome of this preliminary exercise highlighted that multiple issues need to be explored for achieving a scientifically more sound risk assessment and for the development of robust sediment quality criteria for copper, including (i) the use of the assessment factor approach vs. the statistical extrapolation approach, (ii) the importance of bioavailability modifying factors (e.g., organic carbon, acid volatile sulfide), and (iii) the influence of prevailing geochemical (bioavailable) background concentrations on the copper sensitivity of local benthic biota.

c. Question/Issue: Do we have site data to confirm this?

Response: We do not have site specific data for our state that compares a wide variety of bioassays.

d. Question/Issue: Are there species sensitivity curves available that can confirm this?

Response: Unlike terrestrial or marine datasets, very little comprehensive data is available for freshwater bioassays. Where large datasets are available, the species tend to be limited to the standardized species and endpoints, which cannot be used to develop species sensitivity distributions due to limited number of species.

- e. Question/Issue: Are the bioassays appropriate for diverse types of waterbodies (i.e. different sized lakes, streams, intermittent waterbodies)?

Response: These bioassays are appropriate within specified parameters that include pH, grain size, and others listed in the ASTM protocols. The rule will be sufficiently flexible that when sediments are unusual, other assays may be applied as appropriate.

- f. Question/Issue: pH, grain size etc?

Response: These organisms are standardized in part because they are fairly flexible regarding environmental conditions. Standardized assays provide the range of conditions that are acceptable for pH and several other parameters. However, in cases where sediments have unusual characteristics, reference sediments should be found with similar characteristics as a point of comparison. Ecology plans on providing guidance on this topic, but will not include it in the rule.

3. Need some flexibility to request use of different suite of bioassays:

- a. Question/Issue: Make clear which waterbodies they are good predictors

Response: The regional data set used in this effort represents a substantial increase in geographically diverse locations from across the state. However we recognize the fact that not all freshwater environments are represented. For this reason, Ecology will implement the SMS regulatory framework that relies on the biological override where it is suspected that conditions deviate from the norm. It will be important to provide guidance at the time of adopting the freshwater standards that specifically call out the conditions where a site manager can rely on the adopted SQVs or should use a suite of bioassays to characterize the sediments. These could include, but not be limited to, pH, hardness, total organic carbon, and total volatile solids. These parameters and appropriate ranges have not yet been determined.

- b. Question/Issue: List other tests/organisms that could be used in guidance, preferable the rule.

Response: Ecology can allow use of additional species where they are relevant. Guidance will include the ranges of total organic carbon, grain size, pH, and other parameters where standardized tests are appropriate, and where they may be appropriate if suitable reference material is available, and where they may not be appropriate. Other species/tests will not be listed in the rule, in order to provide flexibility, but may be included in the guidance.

Questions from the Science Panel related to the Floating Percentile Method

1. Question/Issue: Why was optimization of false negatives and false positives a goal and why was this important?

Response: This was a policy call by Ecology to develop the Freshwater Sediment Quality Values (SQV) as a tool for *managing risk* to the aquatic environment. The goal was to achieve the greatest accuracy in predicting biological toxicity in the bioassay suite and entailed the choice to avoid both over-predicting or under-predicting toxicity. The goal to develop SQVs that were the most accurate predictor of hits using the biological toxicity tests was based on following the existing sediment standards framework where the biology overrides chemistry, taking into account the following assumptions:

- The Sediment Management Standards (SMS) framework used in the SMS gives priority to the biological tests. The suite of bioassays overrides chemical results.
- The available suite of bioassays serve as the best surrogate available for measuring potential effects from contaminated sediments to the benthic community
- The SMS framework allows some adverse effects within a limited range, bounded by the Sediment Quality Standard (SQS) at the lower level and the Cleanup Screening Level (CSL) at the upper level. (For each of the bioassay endpoints, SQS level is determined by the Minimum Detectable Difference (MDD) and the CSL is established as a 10 or 15 percent increase above that MDD.)
- The chemical SQVs are developed to be the most accurate predictor of when contaminated sediment will cause toxicity in the suite of bioassays.

The advantage in using chemistry over a suite of bioassays to assess a site is the lower cost. This benefit is lost if the chemistry were set so low as to screen all sites as requiring cleanup or further biological testing.

This approach for developing SQVs differs from the goals associated with the development of other guidelines where a lower guideline is established below which there will be a high confidence there is no toxicity (e.g., the Threshold Effects Level where there is a low chance of false negatives but a very high chance for false positives) and a higher guideline above which there is a high confidence of toxicity (e.g., the Probable Effects Level for which there is a low chance of false positives but a high potential for false negatives). These other guidelines serve as a good tool for assessing risk to the benthic community but do not serve well in managing that risk.

2. Question/Issue: How will Ecology avoid missing sites given the policy decision to begin with the 20th percentile level for false negatives?

Response: It is important to note that the selected initial level for false negatives is for each bioassay endpoint and that the FPM method includes combining multiple endpoints from which the lowest is selected as the SQS. This results in a more

conservative SQS or lower overall false negative rate for the proposed SQVs . Further conservatism is inherent in the characterization of a site when each sample station is evaluated individually against the SQVs, increasing the potential for exceeding the standards if elevated chemical concentrations occur at the site.

3. Question/Issue: Co-occurrence and Covariance: The SQV report needs to better define these terms and ensure they are used properly.

Response: Ecology agrees these terms and how they are used in the report needs to be clear. These edits are being made, but were not included in the rewrites presently available. Additional description of how the FPM approach handles covariance will be added to the report.

The primary method this is dealt with is the combining or summing of chemicals with similar modes of action like the PAHs. Early testing of the different methods for combining and reporting these groups of compounds was performed to determine which ones most improved the reliability of the SQVs. This was discussed in Appendix B to the 2010 SQV report and is further discussed in the 2003 Phase II Report by Michelsen

Citation: Michelsen, T. (2003) Phase II Report: Development and Recommendation of SQVs for. Freshwater Sediments in Washington State. Dept of Ecology Publication Number 03-09-088. (<http://www.ecy.wa.gov/biblio/0309088.html>)

4. Question/Issue: Reliability Measures and options for addressing them were identified.

Response: Since the Science Panel last convened, we have met with Burt Shephard, EPA Region 10, Office of Environmental Assessment, and Oregon DEQ NWR Cleanup/Portland Harbor Section, to explore methods employed by the Portland Harbor Superfund team to address reliability measures. The primary issues affecting reliability of any model for developing SQVs are the model's values themselves, the prevalence of hits and no-hits in the data set, and the degree of overlap in concentrations for hit and no-hit toxicity results.

In both the Portland Harbor data and Ecology's regional data (which includes Portland Harbor data) there was a considerable range of overlap for many chemicals and this is not handled well by any of the models in discriminating between hits and no-hits. Prevalence also affects reliability for all models but can be addressed by use of reliability statistics that are not affected by prevalence or can be explicitly adjusted to address prevalence of toxicity. In Portland Harbor, it was found that there were a handful of metrics that met these requirements and these were recommended for use in evaluating reliability of Ecology's SQVs as well. These were recently made available to us and we have the preliminary results that are currently being reviewed. (See attached draft Reliability Statistics paper).

Four statistical statistical methods were recommended:

- Bias

- Odds Ratio
- Hanssen-Kuipers Discriminant
- Post-Test Odds Ratio

The initial outcome indicates the following:

- There is a moderate bias towards over-predicting toxicity at the SQS level and an approximately balanced bias (neither over nor under-predicting) at the CSL level.
 - This is corroborated using the Odds Ratios where "...the SQVs for both SQS and CSL levels have odds ratios of roughly 4:1, suggesting an 80% likelihood of exceeding the biological standards using the interpretive criteria in the rule given an SQV exceedance. These are still quite reasonable odds, in line with the policy goals used to calculate the guideline (80% overall accuracy, maximum of 20% false negatives and 20% false positives)."
 - The Hans-Kuipers Discriminant measure provides a goodness of fit measure similar to an r-squared value that is best when there is close agreement between prediction and outcome. This was found to only indicate moderate agreement which is likely the result of using a combined or pooled data set rather than the actual model output (as for a single endpoint).
 - The final Post-Test Odds Ratio is sample dependent and the example is still being worked on at this time.
- a. Question/Issue: Comparisons of reliability for other SQVs is comparing apples to oranges since these were developed using different data, different species and toxicity determinations and methods for establishing standards.

Response: The purpose behind examining reliability for different SQVs was to assess how well the FPM and other SQV models perform in predicting toxicity for our region. Ecology recognizes the selection of different species and toxicity endpoints, the regional effects on hydrologic and geologic influences and other factors all contribute to how a particular model or set of SQVs will predict toxicity. Trying to resolve the actual basis for differences in reliability outcomes for different SQVs goes beyond what was intended.

- b. Question/Issue: How predictive is the chemistry when considering the wide variations in site biology and water chemistry?

Response: See response to Bioassay Question 3.a. above.

- c. Question/Issue: How does the hit/no-hit distribution in the regional data set compare to that used for other SQVs and how does that affect reliability?

Response: See the response to reliability issues, above.