

**To:** RSET SQG Subcommittee

**Date:** 7 December 2006

**From:** Mike Poulsen  
Jennifer Peterson

**Subject:** Evaluation of Reliability of Potential Freshwater Sediment  
Screening Values

In developing Section 7 of the Sediment Evaluation Framework (SEF), the Regional Sediment Evaluation Team (RSET) Sediment Quality Guidelines (SQG) subcommittee discussed some of the uncertainties associated with the draft method. We also discussed reliability measures associated with the proposed values. Given the time constraints with preparing the September 2006 draft SEF, a discussion of the reliability measures was not included in Section 7, with the expectation that information on reliability would be included later in an appendix (Evaluation of Reliability of Proposed Freshwater Sediment Quality Guidelines). The purpose of this memorandum is to present information on reliability measures, and provide material that could be included in the proposed appendix.

### **Tiered Screening Levels Based on Toxicity**

As presented in Section 7, RSET has proposed to use draft un-promulgated sediment quality guidelines (SQGs) developed by Dr. Teresa Michelsen for Washington state in 2002. Teresa used a floating percentile method to derive the draft SQGs.

Two screening levels are being proposed, based on different definitions of toxicity ("hit").

- SL1 values are based on a 10% difference (e.g., in mortality or another toxic endpoint such as growth) in a bioassay result compared with the control (a "clean" sample).
- SL2 values are based on a 25% difference from control, indicating a higher degree of toxicity.

In the traditional three tier system of screening, SL1 values will be used to screen out sediment samples as non-toxic. SL2 values will be used to screen in samples as toxic. Samples with concentrations between SL1 and SL2 values will require additional evaluation and other lines of evidence.

### **Evaluating the Reliability of Screening Levels**

Another consideration in the determination of the screening values is how accurate the methodology predicts proper results. Two key reliability measures are the following:

- False negative – the percentage of known toxic samples that are incorrectly screened out using specified screening values
- False positive – the percentage of known non-toxic samples that are incorrectly screened in

Table 1 shows the definitions of all the reliability measures, as well as the results of Teresa's calculations. The SQG subcommittee recommended using screening values developed from a false negative rate of 15%. Figure 1 shows a representation of the results for SL1 using the

false negative rate of 15%. Given our definition of a hit, all samples must be placed in one of four bins, either a correctly predicted hit, a correctly predicted no-hit, a hit incorrectly predicted as a no-hit, or a no-hit incorrectly predicted as a hit. In Figure 1, the number of correctly predicted hits is 34. Dividing by the total number of hits (40) gives a sensitivity of 85%. The corresponding false negative rate is 15% (= 100% - 85%). In other words, using the SL1 screening values, we will miss only 15% of the samples known to be toxic. This is a reasonably good result, and this measure should be a primary focus of the agencies.

However, RSET should also be interested in how reliably we predict no-hits (measured by predicted-no-hit efficiency). Based on our definition of a hit and the proposed screening criteria, Figure 1 shows that we are 67% confident that a sample predicted to be a no-hit at the SL1 screening level will in fact be non-toxic if we were to conduct a bioassay. Stated another way, in the existing dataset, one third of the samples predicted as no-hit were toxic in a bioassay.

The above results give different perspectives on the reliability of the screening values. On one hand, we are reasonably confident that we will screen in known toxic samples at the SL1 level. On the other hand, for samples screened out as non-toxic, there is still a good chance that they may still be toxic.

RSET has not established criteria for making a decision regarding the acceptability of these reliability results. We expect that federal and state managers will likely accept a false negative rate of 15%. It is less clear how concerned managers will be about a false-predicted-no-hit rate of 33%, particularly if the screening values are used as a sole line of evidence. Given this concern, the SQG subcommittee agreed to state in the draft guidance that regulatory agencies may require additional evaluations (possibly including bioassays) even if concentrations of chemicals in sediment are below SL1 screening values.

Reliability results differ with the selected level of toxicity (SL1 or SL2). At the level of toxicity used to develop the SL2 screening values, the predicted-no-hit rate increases from 67% (for SL1) to about 84% (see Table 1), with a correspondingly lower false-predicted-no-hit rate (16%). For a sample with concentrations below SL2 screening levels, there is only a 1/6 chance of the sediment being toxic at the SL2 higher level of toxicity.

In addition, the regulated community will rightly be concerned about the other reliability measures (e.g., false positives and false predicted hits). One of the great benefits of the floating percentile method is the ability to optimize screening values by reducing false positives for a given false negative rate.

The RSET SQG subcommittee discussed the possibility of using more conservative screening values, such as threshold effect levels (TELs), for the lower screening values. TEL values have more optimal false negative and false-predicted-no-hit rates for a lower screen. The reliability estimates for various screening approaches are shown in Table 2. Actual chemical concentration values for the different screening approaches are shown in Table 3.

Some subcommittee members considered the false positive rates too high, and were concerned that very few sediment areas will be screened out if values such as TELs are used for the lower

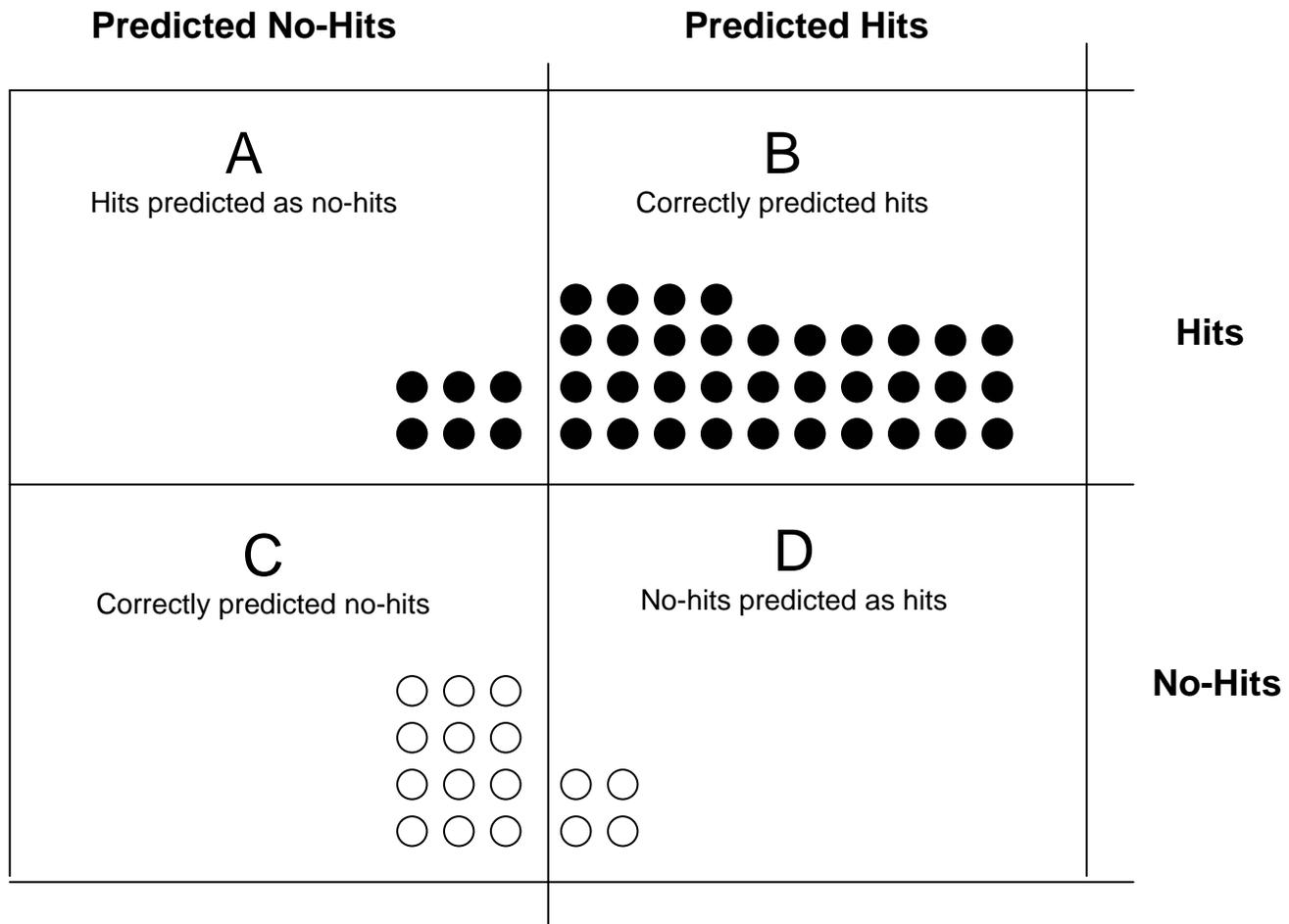
---

screen. RSET has not established explicit criteria for the acceptability of these rates. The compromise was to allow states and other regulatory agencies the option of requiring additional evaluation (including bioassays) for samples with concentrations below SL1 screening values. As the floating percentile screening method is validated, it would be useful if RSET established acceptable criteria for the various reliability measures. The reliability measures are general, and should be examined for any set of screening criteria, regardless of the development method.

### **Uncertainty in the Data Used for Model Development**

In addition to the specific numeric measures of reliability calculated for the model, we are concerned that the current model is based on a limited dataset, primarily from western Washington and western Oregon. Currently, the model is based on 25 sites and 229 samples. This is a limited dataset if the goal is to develop a regional predictive dataset. For example, the bioavailability of PAHs may be very low in many of the samples included in the dataset due to the source of the material. These sites are not likely to be representative of other sites, and the result may be low predictive power for PAH sites not included in model development. We consider it necessary to incorporate additional data to add robustness to the model. We also consider it necessary to validate the model using data from a variety of environments, and using data separate from the data used to develop the screening values.

**Figure 1.**  
**Reliability Measures of Proposed SL1 Screening Criteria**



- Adverse effects observed (hit)
- No adverse effects observed (no-hit)

Sensitivity =  $B / (A + B) = 0.85$   
 False Negatives =  $A / (A + B) = 0.15$

Predicted-Hit Efficiency =  $B / (B + D) = 0.89$   
 False Predicted Hits =  $D / (B + D) = 0.11$

Efficiency =  $C / (C + D) = 0.75$   
 False Positives =  $D / (C + D) = 0.25$

Predicted-No-Hit Efficiency =  $C / (A + C) = 0.67$   
 False Predicted No-Hits =  $A / (A + C) = 0.33$

**Table 1**  
**Reliability Estimates for Proposed Freshwater Sediment Screening Values**  
**in Draft SEF**

Reliability Measure	Definition	Percentage (%)	
		Screening Level: SL1	SL2
Sensitivity (Hit Efficiency)	Percentage of known toxic samples that are correctly screened in	84	85
False Negative	Percentage of known toxic samples that are incorrectly screened out	16	15
(No-Hit) Efficiency	Percentage of known non-toxic samples that are correctly screened out	75	75
False Positive	Percentage of known non-toxic samples that are incorrectly screened in	25	25
Predicted-Hit Efficiency	Percentage of screened-in samples that are toxic	88	77
False Predicted Hit	Percentage of screened-in samples that are non-toxic	12	23
Predicted-No-Hit Efficiency	Percentage of screened-out samples that are non-toxic	67	84
False Predicted-No-Hit	Percentage of screened-out samples that are toxic	33	16

Note:  
 See Figure 1 for a graphical presentation of the reliability measures for SL1.



**Table 3**  
**Comparison of Proposed RSET Freshwater Sediment Screening Values**  
**With Other Screening Values**

Chemical	Proposed RSET Screening Levels <sup>a</sup>		Other Freshwater Sediment Values <sup>b,c</sup>				
	SL1	SL2	TEL	TEC	PEL	PEC	AET
<b>Metals (mg/kg)</b>							
Antimony	0.4	0.6					64
Arsenic	20	51	5.9	9.8	17	33	40
Cadmium	0.6	1	0.6	0.99	3.5	4.5	7.6
Chromium	95	100	37	43	90	110	280
Copper	80	830	36	32	200	150	840
Lead	335	430	35	36	91	130	260
Mercury	0.5	0.75	0.17	0.18	0.49	1.1	0.56
Nickel	60	70	18	23	36	49	46
Silver	2	2.5					4.5
Zinc	140	160	120	120	320	460	520
Tributyltin	75	75					
<b>SVOCs (ug/kg)</b>							
Total PCBs	60	120	34	60	280	680	21
DEHP	230	320					750
Butylbenzylphthalate	260	370					
Di-n-butylphthalate							
Dibenzofuran	400	440					32,000
<b>Pesticides (ug/kg)</b>							
Total DDTs			1.2	5.3	4.8	570	
<b>PAHs (ug/kg)</b>							
Total LPAH	6,600	9,200					74,000
Total HPAH	31,000	54,800					91,000
Total PAHs				1,600		23,000	170,000
Acenaphthene	1,060	1,320	6.7		89		4,100
Acenaphthylene	470	640	5.9		130		2,200
Anthracene	1,200	1,580	47	57	250	850	2,800
Benz[a]anthracene	4,260	5,800	32	110	390	1,100	7,700
Benzo[a]pyrene	3,300	4,810	32	150	780	1,500	11,000
Benzo[g,h,i]perylene	4,020	5,200					1,400
Chrysene	5,940	6,400					
Dibenz[a,h]anthracene	800	840	6.2	33	140		230
Fluoranthene	11,000	15,000	110	420	2,400	2,200	21,000
Fluorene	1,000	3,000	21	77	140	540	4,200
Naphthalene	500	1,310	35	180	390	560	46,000
Phenanthrene	6,100	7,600	42	200	520	1,200	15,000
Pyrene	8,800	16,000	53	200	880	200	23,000

Notes:

- a) Draft Sediment Evaluation Framework, Sept. 2005, Table 7-1.
- b) Development of Freshwater Sediment Quality Values for Use in Washington State, Phase I Task 6 Report, Sept. 2002, Appendix H.
- c) SL1 = Screening Level 1  
 TEL = Threshold Effects Level  
 PEL = Probable Effects Level  
 AET = Apparent Effects Threshold
- SL2 = Screening Level 2  
 TEC = Threshold Effects Concentration  
 PEC = Probably Effects Concentration