Tables

Table 1. Ecology's quality assurance and adverse effects (toxicity) levels for biological tests, as presented in Table VIII.

Test	QA Control	SCOs	CSLs
Chironomus dilutus 10-day mortality	C ≤ 30% [a]	T - C > 20%	T - C > 30%
Chironomus dilutus 10-day growth	CF ≥ 0.48 mg/ind	T/C < 80%	T/C < 70%
Hyalella azteca 10-day mortality	C ≤ 20% [a]	T - C > 15%	T - C > 25%
Hyalella azteca 28-day mortality	C ≤ 20% [a]	T - C > 10%	T - C > 25%
Hyalella azteca 28-day growth	CF ≥ 0.15 mg/ind	T/C < 75%	T/C < 60%

QA = quality assurance; SCO = sediment cleanup objective; CSL = cleanup screening level; C = control; CF = control final; T = test sample.

[a] These control mortality limits are currently in the process of being reviewed by ASTM and may be lowered in the next few years (Ingersoll *et al* . 2008; as cited in WDOE 2011).

Table 2. Comparison of methods used to designate sediments as toxic or not toxic to the midge (*Chironomus dilutus*) from samples collected from the Upper Columbia River in 2005.

Station	Batch	10-day Percent Survival					10-day Growth (weight)				
		Effect Value (%)	Control-adjusted Value (%)	Reference Envelope	Draft SMS SCO	Draft SMS CSL	Effect Value (mg)	Control-adjusted Value (%)	Reference Envelope	Draft SMS SCO	Draft SMS CSL
Control	1	83.8	100	--	--	--	1.51	100	--	--	--
Control	2	88.8	100	--	--	--	1.97	100	--	--	--
RM727A1(X1)	2	92.5	104	NT	NT	NT	1.74	88.1	T	NT	NT
RM729A1(X1)	2	90	101	NT	NT	NT	2.01	102	NT	NT	NT
RM628A1(X1)	2	87.5	98.5	NT	NT	NT	1.93	97.7	T	NT	NT
RM742A1(X1)	1	82.5	98.4	NT	NT	NT	1.18	78.3	T	NT	NT
RM743A1(X1)	1	82.5	98.4	NT	NT	NT	1.60	106	T	NT	NT
RM641A1(X1)	2	86.3	97.1	NT	NT	NT	1.93	97.8	T	NT	NT
RM743A2(X3)	1	80	95.5	NT	NT	NT	1.43	94.5	T	NT	NT
RM616A1(X3)	2	83.8	94.3	NT	NT	NT	2.22	113	NT	T	NT
RM723A1(X1)	2	83.8	94.3	NT	NT	NT	2.14	109	NT	NT	NT
RM723A2(X3)	2	83.8	94.3	NT	NT	NT	1.96	99.3	NT	NT	NT
RM733A1(X1)	2	83.8	94.3	NT	NT	NT	1.76	89.5	T	NT	NT
RM730A1	2	82.5	92.9	NT	NT	NT	1.96	99.7	NT	NT	NT
RM737A1(X3)	2	82.5	92.9	NT	NT	NT	1.47	74.4	T	T	T
RM605A2(X8)	2	81.3	91.5	NT	NT	NT	1.85	94	T	NT	NT
RM661A1(X1)	2	81.3	91.5	NT	NT	NT	1.84	93.3	T	NT	NT
RM734A1	2	81.3	91.5	NT	NT	NT	1.61	81.8	T	NT	NT
RM736A1(X1)	2	81.3	91.5	NT	NT	NT	1.94	98.5	NT	NT	NT
RM744A2(X3)	1	76.3	91	NT	NT	NT	1.31	86.8	T	NT	NT
RM634A1(X1)	2	80	90.1	NT	NT	NT	1.92	97.3	T	NT	NT
RM658A1(X3)	2	80	90.1	NT	NT	NT	1.78	90.3	T	NT	NT
RM740A1(X1)	1	75	89.5	NT	NT	NT	2.08	137	NT	NT	NT
RM687A1	1	73.8	88	NT	NT	NT	1.62	107	T	NT	NT
RM742A2(X5)	1	73.8	88	NT	NT	NT	1.31	86.5	T	NT	NT
RM605A1(X1)	2	77.5	87.3	NT	NT	NT	1.81	91.9	T	NT	NT
RM622A1(X3)	2	77.5	87.3	NT	NT	NT	1.82	92.4	T	NT	NT
RM637A1(X1)	2	77.5	87.3	NT	NT	NT	1.79	90.9	T	NT	NT

Table 2. Comparison of methods used to designate sediments as toxic or not toxic to the midge (*Chironomus dilutus*) from samples collected from the Upper Columbia River in 2005.

Station	Batch	10-day Percent Survival					10-day Growth (weight)				
		Effect Value (%)	Control-adjusted Value (%)	Reference Envelope	Draft SMS SCO	Draft SMS CSL	Effect Value (mg)	Control-adjusted Value (%)	Reference Envelope	Draft SMS SCO	Draft SMS CSL
RM686A1(X3)	1	72.5	86.5	NT	NT	NT	1.83	121	T	NT	NT
RM692A1(X1)	1	72.5	86.5	NT	NT	NT	1.80	119	T	NT	NT
RM739A1(X3)	1	72.5	86.5	NT	NT	NT	2.04	135	NT	NT	NT
RM678A1(X1)	1	71.3	85	NT	NT	NT	1.93	128	T	NT	NT
RM713A1(X3)	2	75	84.5	NT	NT	NT	2.23	113	NT	NT	NT
RM708A1(X3)	1	70	83.5	NT	NT	NT	1.65	109	T	NT	NT
RM706A1(X1)	1	68.8	82	NT	NT	NT	1.69	112	T	NT	NT
RM606A1(X3)	2	72.5	81.6	NT	NT	NT	2.04	104	NT	T	T
RM698A1(X1)	1	67.5	80.5	NT	NT	NT	1.75	116	T	NT	NT
RM738A1(X3)	1	67.5	80.5	NT	NT	NT	1.14	75.7	T	NT	NT
RM741A1(X3)	1	67.5	80.5	NT	NT	NT	2.18	144	NT	NT	NT
RM603A1(X1)	2	71.3	80.2	NT	NT	NT	1.90	96.6	T	NT	NT
RM644A1(X3)	2	70	78.8	T	NT	NT	1.82	92.4	T	NT	NT
RM724A2(X3)	1	65	77.6	NT	NT	NT	2.44	162	NT	NT	NT
RM704A1(X1)	1	62.5	74.6	NT	T	NT	2.02	134	NT	NT	NT
RM642A1(X1)	2	66.3	74.6	T	NT	NT	1.97	100	NT	NT	NT
RM744A1(X1)	1	61.3	73.1	T	T	NT	1.98	131	NT	NT	NT
RM640A1(X3)	2	60	67.6	T	T	NT	2.52	128	NT	NT	NT
RM724A1(X1)	1	56.3	67.1	T	T	NT	2.13	141	NT	NT	NT
RM706A2(X7)	1	55	65.6	T	T	NT	2.05	135	NT	NT	NT
RM689A1(X3)	1	50	59.7	T	T	T	2.37	157	NT	NT	NT
RM676A1(X3)	1	46.3	55.2	T	T	T	2.10	139	NT	NT	NT
RM677A1(X3)	1	42.5	50.7	T	T	T	1.99	132	NT	NT	NT
RM680A1(X1)	1	38.8	46.2	T	T	T	2.17	143	NT	NT	NT

SMS = sediment management standards; SCO = sediment cleanup objective; CSL = cleanup screening level; T = toxic; NT = not toxic

Table 3. Comparison of methods used to designate sediments as toxic or not toxic to the amphipod (*Hyalella azteca*) from samples collected from the Upper Columbia River in 2005.

Station	Batch	28-day Percent Survival					28-day Growth (weight)				
		Effect Value (%)	Control-adjusted Value (%)	Reference Envelope	Draft SMS SCO	Draft SMS CSL	Effect Value (mg)	Control-adjusted Value (%)	Reference Envelope	Draft SMS SCO	Draft SMS CSL
Control	1	96.3	100	--	--	--	0.41	100	--	--	--
Control	2	97.5	100	--	--	--	0.38	100	--	--	--
RM724A1(X1)	1	98.8	103	NT	NT	NT	0.371	90.5	NT	NT	NT
RM713A1(X3)	2	100	103	NT	NT	NT	0.334	87.9	NT	NT	NT
RM678A1(X1)	1	97.5	101	NT	NT	NT	0.334	81.5	NT	T	NT
RM740A1(X1)	1	97.5	101	NT	NT	NT	0.515	126	NT	NT	NT
RM658A1(X3)	2	98.8	101	NT	NT	NT	0.412	108	NT	NT	NT
RM661A1(X1)	2	97.5	100	NT	NT	NT	0.345	90.8	NT	NT	NT
RM727A1(X1)	2	97.5	100	NT	NT	NT	0.412	108	NT	NT	NT
RM680A1(X1)	1	96.3	99.9	NT	NT	NT	0.326	79.5	NT	NT	NT
RM698A1(X1)	1	96.3	99.9	NT	NT	NT	0.285	69.5	T	T	NT
RM704A1(X1)	1	96.3	99.9	NT	NT	NT	0.383	93.4	NT	NT	NT
RM692A1(X1)	1	95	98.7	NT	NT	NT	0.412	100	NT	NT	NT
RM706A1(X1)	1	95	98.7	NT	NT	NT	0.312	76.1	NT	NT	NT
RM706A2(X7)	1	95	98.7	NT	NT	NT	0.335	81.7	NT	NT	NT
RM724A2(X3)	1	95	98.7	NT	NT	NT	0.642	157	NT	NT	NT
RM742A2(X5)	1	95	98.7	NT	NT	NT	0.32	78	NT	NT	NT
RM603A1(X1)	2	96.3	98.7	NT	NT	NT	0.324	85.3	NT	NT	NT
RM616A1(X3)	2	96.3	98.7	NT	NT	NT	0.484	127	NT	NT	NT
RM637A1(X1)	2	96.3	98.7	NT	NT	NT	0.451	119	NT	NT	NT
RM640A1(X3)	2	96.3	98.7	NT	NT	NT	0.401	106	NT	NT	NT
RM642A1(X1)	2	96.3	98.7	NT	NT	NT	0.301	79.2	NT	NT	NT
RM723A1(X1)	2	96.3	98.7	NT	NT	NT	0.652	172	NT	NT	NT
RM686A1(X3)	1	93.8	97.4	T	NT	NT	0.592	144	NT	NT	NT
RM687A1	1	93.8	97.4	T	NT	NT	0.268	65.4	T	NT	NT
RM689A1(X3)	1	93.8	97.4	T	NT	NT	0.368	89.8	NT	NT	NT
RM622A1(X3)	2	95	97.4	NT	NT	NT	0.516	136	NT	NT	NT
RM641A1(X1)	2	95	97.4	NT	NT	NT	0.348	91.6	NT	NT	NT

Table 3. Comparison of methods used to designate sediments as toxic or not toxic to the amphipod (*Hyalella azteca*) from samples collected from the Upper Columbia River in 2005.

Station	Batch	28-day Percent Survival					28-day Growth (weight)				
		Effect Value (%)	Control-adjusted Value (%)	Reference Envelope	Draft SMS SCO	Draft SMS CSL	Effect Value (mg)	Control-adjusted Value (%)	Reference Envelope	Draft SMS SCO	Draft SMS CSL
RM723A2(X3)	2	95	97.4	NT	NT	NT	0.423	111	NT	NT	NT
RM605A2(X8)	2	93.8	96.2	T	NT	NT	0.334	87.9	NT	NT	NT
RM606A1(X3)	2	93.8	96.2	T	NT	NT	0.421	111	NT	NT	NT
RM676A1(X3)	1	92.5	96.1	T	NT	NT	0.347	84.6	NT	NT	NT
RM708A1(X3)	1	92.5	96.1	T	NT	NT	0.339	82.7	NT	NT	NT
RM605A1(X1)	2	92.5	94.9	T	NT	NT	0.486	128	NT	NT	NT
RM644A1(X3)	2	92.5	94.9	T	NT	NT	0.341	89.7	NT	NT	NT
RM729A1(X1)	2	92.5	94.9	T	NT	NT	0.449	118	NT	NT	NT
RM739A1(X3)	1	91.3	94.8	T	NT	NT	0.461	112	NT	NT	NT
RM743A1(X1)	1	91.3	94.8	T	NT	NT	0.486	119	NT	NT	NT
RM634A1(X1)	2	91.3	93.6	T	NT	NT	0.371	97.7	NT	NT	NT
RM733A1(X1)	2	91.3	93.6	T	NT	NT	0.499	131	NT	NT	NT
RM677A1(X3)	1	90	93.5	T	NT	NT	0.276	67.3	T	T	NT
RM737A1(X3)	2	90	92.3	T	NT	NT	0.194	51.1	T	T	T
RM742A1(X1)	1	88.8	92.2	T	NT	NT	0.273	66.6	T	T	NT
RM736A1(X1)	2	88.8	91.0	T	NT	NT	0.336	88.4	NT	NT	NT
RM738A1(X3)	1	86.3	89.6	T	NT	NT	0.178	43.4	T	T	T
RM730A1	2	86.3	88.5	T	T	NT	0.336	88.4	NT	NT	NT
RM734A1	2	86.3	88.5	T	T	NT	0.227	59.7	T	T	T
RM744A1(X1)	1	83.8	87	T	T	NT	0.349	85.1	NT	NT	NT
RM628A1(X1)	2	83.8	85.9	T	NT	NT	0.524	138	NT	NT	NT
RM743A2(X3)	1	81.3	84.4	T	T	NT	0.321	78.3	NT	NT	NT
RM741A1(X3)	1	80	83.1	T	T	NT	0.46	112	NT	NT	NT
RM744A2(X3)	1	75	77.9	T	T	NT	0.166	40.5	T	T	T

SMS = sediment management standards; SCO = sediment cleanup objective; CSL = cleanup screening level; T = toxic; NT = not toxic

Table 4. Comparison showing the number of stations designated toxic from the Upper Columbia River 2005 study according to each method.

Test	Reference Envelope	Draft SMS SCO	Draft SMS CSL
Chironomus dilutus 10-day Percent survival	10 of 56 (18%)	9 of 50 (18%)	4 of 50 (8%)
Chironomus dilutus 10-day Growth (weight)	28 of 56 (50%)	3 of 50 (6%)	2 of 50 (4%)
Hyalella azteca 28-day Percent survival	26 of 56 (46%)	6 of 50 (12%)	0 of 50 (0%)
Hyalella azteca 28-day Growth (weight)	8 of 56 (14%)	8 of 50 (16%)	4 of 50 (8%)

SMS = sediment management standards; SCO = sediment cleanup objective; CSL = cleanup screening level.

Table 5. Comparison of sediment cleanup objectives (SCOs) to sediment quality guidelines in freshwater ecosystems that reflect TECs (i.e., below which harmful effects are unlikely to be observed).

Substance	*Threshold Effect Concentrations*							
	TEL	**LEL**	**MET**	**ERL**	**TEL-HA28**	**SQAL**	**Consensus-Based TEC**	**SCO**
Metals (mg/kg DW)								
Arsenic	5.9	6	7	33	11	NG	9.79	14
Cadmium	0.596	0.6	0.9	5	0.58	NG	0.99	2.1
Chromium	37.3	26	55	80	36	NG	43.4	72
Copper	35.7	16	28	70	28	NG	31.6	400
Lead	35	31	42	35	37	NG	35.8	360
Mercury	0.174	0.2	0.2	0.15	NG	NG	0.18	0.66
Nickel	18	16	35	30	20	NG	22.7	26
Zinc	123	120	150	120	98	NG	121	3200
Polycyclic Aromatic Hydrocarbons (PAH; µg/kg DW)								
Anthracene	NG	220	NG	85	10	NG	57.2	NG
Fluorene	NG	190	NG	35	10	540	77.4	NG
Naphthalene	NG	NG	400	340	15	470	176	NG
Phenanthrene	41.9	560	400	225	19	1800	204	NG
Benz[a]anthracene	31.7	320	400	230	16	NG	108	NG
Benzo(a)pyrene	31.9	370	500	400	32	NG	150	NG
Chrysene	57.1	340	600	400	27	NG	166	NG
Dibenz[a,h]anthracene	NG	60	NG	60	10	NG	33.0	NG
Fluoranthene	111	750	600	600	31	6200	423	NG
Pyrene	53	490	700	350	44	NG	195	NG
Total PAHs	NG	4000	NG	4000	260	NG	1610	17000
Polychlorinated Biphenyls (PCB; µg/kg DW)								
Total PCBs	34.1	70	200	50	32	NG	59.8	110
Organochlorine Pesticides (µg/kg DW)								
Chlordane	4.5	7	7	0.5	NG	NG	3.24	NG
Dieldrin	2.85	2	2	0.02	NG	110	1.90	4.9

Table 5. Comparison of sediment cleanup objectives (SCOs) to sediment quality guidelines in freshwater ecosystems that reflect TECs (i.e., below which harmful effects are unlikely to be observed).

Substance	*Threshold Effect Concentrations*							
	TEL	**LEL**	**MET**	**ERL**	**TEL-HA28**	**SQAL**	**Consensus-Based TEC**	**SCO**
Organochlorine Pesticides (µg/kg DW; cont.)								
Sum DDD	3.54	8	10	2	NG	NG	4.88	310
Sum DDE	1.42	5	7	2	NG	NG	3.16	21
Sum DDT	NG	8	9	1	NG	NG	4.16	100
Total DDTs	7	7	NG	3	NG	NG	5.28	NG
Endrin	2.67	3	8	0.02	NG	42	2.22	8.5 [1]
Heptachlor epoxide	0.6	5	5	NG	NG	NG	2.47	NG
Lindane (gamma-BHC)	0.94	3	3	NG	NG	3.7	2.37	NG

TEL = threshold effect level, dry weight (Smith *et al*. 1996); LEL = lowest effect level, dry weight (Persaud *et al*. 1993); MET = minimal effect threshold, dry weight (EC & MENVIQ 1992); ERL = effects range low, dry weight (Long and Morgan 1991); TEL-HA28 = threshold effect level for *Hyalella azteca,* 28-day test, dry weight (USEPA 1996); SQAL = sediment quality advisory levels, dry weight at 1% OC (USEPA 1997); TEC = threshold effect concentration (MacDonald *et al*. 2000); SCO = sediment cleanup objective, dry weight; NG = no guideline; OC = organic carbon; DDD = dichlorodiphenyldichloroethane; DDE = dichlorodiphenyldichloroethylene; DDT = dichlorodiphenyltrichloroethane

[1] Guideline for endrin ketone

Table 6. Comparison of cleanup screening levels to sediment quality guidelines in freshwater ecosystems that reflect PECs (i.e., above which harmful effects are likely to be observed).

Substance	*Probable Effect Concentrations*						
	PEL	**SEL**	**TET**	**ERM**	**PEL-HA28**	**Consensus-Based PEC**	**CSL**
Metals (mg/kg DW)							
Arsenic	17	33	17	85	48	33.0	120
Cadmium	3.53	10	3	9	3.2	4.98	5.4
Chromium	90	110	100	145	120	111	88
Copper	197	110	86	390	100	149	1200
Lead	91.3	250	170	110	82	128	> 1300
Mercury	0.486	2	1	1.3	NG	1.06	0.8
Nickel	36	75	61	50	33	48.6	110
Zinc	315	820	540	270	540	459	> 4200
Polycyclic Aromatic Hydrocarbons (PAH; µg/kg DW)							
Anthracene	NG	3700	NG	960	170	845	NG
Fluorene	NG	1600	NG	640	150	536	NG
Naphthalene	NG	NG	600	2100	140	561	NG
Phenanthrene	515	9500	800	1380	410	1170	NG
Benz[a]anthracene	385	14800	500	1600	280	1050	NG
Benzo(a)pyrene	782	14400	700	2500	320	1450	NG
Chrysene	862	4600	800	2800	410	1290	NG
Fluoranthene	2355	10200	2000	3600	320	2230	NG
Pyrene	875	8500	1000	2200	490	1520	NG
Total PAHs	NG	100000	NG	35000	3400	22800	30000
Polychlorinated Biphenyls (PCB; µg/kg DW)							
Total PCBs	277	5300	1000	400	240	676	2500
Organochlorine Pesticides (µg/kg DW)							
Chlordane	8.9	60	30	6	NG	17.6	NG
Dieldrin	6.67	910	300	8	NG	61.8	9.3

Table 6. Comparison of cleanup screening levels to sediment quality guidelines in freshwater ecosystems that reflect PECs (i.e., above which harmful effects are likely to be observed).

Substance	*Probable Effect Concentrations*						
	PEL	**SEL**	**TET**	**ERM**	**PEL-HA28**	**Consensus-Based PEC**	**CSL**
Organochlorine Pesticides (µg/kg DW; cont.)							
Sum DDD	8.51	60	60	20	NG	28.0	860
Sum DDE	6.75	190	50	15	NG	31.3	33
Sum DDT	NG	710	50	7	NG	62.9	8100
Total DDTs	4450	120	NG	350	NG	572	NG
Endrin	62.4	1300	500	45	NG	207	NG
Heptachlor Epoxide	2.74	50	30	NG	NG	16.0	NG
Lindane (gamma-BHC)	1.38	10	9	NG	NG	4.99	NG

PEL = probable effect level, dry weight (Smith *et al.* 1996); SEL = severe effect level, dry weight (Persaud *et al* . 1993); TET = toxic effect threshold, dry weight (EC & MENVIQ 1992); ERM = effects range median, dry weight (Long and Morgan 1991); PEL-HA28 = probable effect level for *Hyalella azteca* ; 28 day test; dry weight (USEPA 1996); PEC = probable effect concentration (MacDonald *et al.* 2000); CSL = cleanup screening level, dry weight; NG = no guideline; OC = organic carbon; NG = no guideline; OC = organic carbon; DDD = dichlorodiphenyldichloroethane; DDE = dichlorodiphenyldichloroethylene; DDT = dichlorodiphenyltrichloroethane

Table 7. Comparison of sediment cleanup objectives and cleanup screening levels to sediment quality standards in freshwater ecosystems in Washington State.

Substance	*Threshold Effect Concentrations*			*Probable Effect Concentrations*	
	Confederated Tribes of the Colville Reservation	Spokane Tribe of Indians	SCO	Confederated Tribes of the Colville Reservation	CSL
Metals (mg/kg DW)					
Arsenic	9.79	9.79	14	33.0	120
Cadmium	0.99	0.99	2.1	4.98	5.4
Chromium	43.4	43.4	72	111	88
Copper	31.6	31.6	400	149	1200
Lead	35.8	35.8	360	128	> 1300
Mercury	0.18	0.18	0.66	1.06	0.8
Nickel	22.7	22.7	26	48.6	110
Zinc	121	121	3200	459	> 4200
Polycyclic Aromatic Hydrocarbons (PAH; µg/kg DW)					
Anthracene	57.2	57.2	NG	845	NG
Fluorene	77.4	77.4	NG	536	NG
Naphthalene	176	176	NG	561	NG
Phenanthrene	204	204	NG	1170	NG
Benz[a]anthracene	108	108	NG	1050	NG
Benzo(a)pyrene	150	150	NG	1450	NG
Chrysene	166	166	NG	1290	NG
Dibenz[a,h]anthracene	33.0	33.0	NG	NG	NG
Fluoranthene	423	423	NG	2230	NG
Pyrene	195	195	NG	1520	NG
Total PAHs	1610	1610	17000	22800	30000
Polychlorinated Biphenyls (PCB; µg/kg DW)					
Total PCBs	59.8	59.8	110	676	2500
Organochlorine Pesticides (µg/kg DW)					
Chlordane	3.24	3.24	NG	17.6	NG
Dieldrin	1.90	1.90	4.9	61.8	9.3
Sum DDD	4.88	4.88	310	28.0	860
Sum DDE	3.16	3.16	21	31.3	33
Sum DDT	4.16	4.16	100	62.9	8100
Total DDTs	5.28	5.28	NG	572	NG
Endrin	2.22	2.22	8.5 [1]	207	NG
Heptachlor epoxide	2.47	2.47	NG	16.0	NG
Lindane (gamma-BHC)	2.37	2.37	NG	4.99	NG

NG = no guideline; SCO = sediment cleanup objective, dry weight; CSL = cleanup screening level, dry weight;
DDD = dichlorodiphenyldichloroethane; DDE = dichlorodiphenyldichloroethylene; DDT = dichlorodiphenyltrichloroethane
[1] Guideline for endrin ketone.

Table 8. Summary of methods used to designate sediments collected in Washington State as toxic or not toxic.

Study	Toxicity Test	Number of Stations	Comparison Method Used for Designating Toxicity	Reference
Upper Columbia River				
Schut and Stefanoff (2007)	10-d *Chironomus dilutus* WST (S, G, B)	56	Reference Envelope	MacDonald *et al.* (2012)
	28-d *Hyalella azteca* WST (S, G, B)	56	Reference Envelope	MacDonald *et al.* (2012)
Besser *et al.* (2008)	12-d *Chironomus dilutus* WST (S, G, B)	8	Reference Envelope	MacDonald *et al.* (2012)
	28-d *Hyalella azteca* WST (S, G, B)	8	Reference Envelope	MacDonald *et al.* (2012)
Bortleson *et al.* (1994)	7-d *Hyalella azteca* WST (S)	19	Negative Control ($\alpha = 0.05$)	Bortleson *et al.* (1994)
Era and Serdar (2001)	20-d *Chironomus dilutus* WST (S, G)	9	Negative Control ($\alpha = 0.05$)	Era and Serdar (2001)
	10-d *Hyalella azteca* WST (S)	9	Negative Control ($\alpha = 0.05$)	Era and Serdar (2001)
Johnson (1991)	10-d *Hyalella azteca* WST (S)	6	Negative Control ($\alpha = 0.05$)	Unpublished analysis by MESL
Washington State				
Johnson and Norton (2001)	20-d *Chironomus dilutus* WST (S, G)	8	Negative Control ($\alpha = 0.05$)	Johnson and Norton (2001)
	28-d *Hyalella azteca* WST (S, G, B)	8	Negative Control ($\alpha = 0.05$)	Johnson and Norton (2001)
Johnson and Plotnkioff (2000)	10-d *Hyalella azteca* WST (S)	4	Negative Control ($\alpha = 0.05$)	Johnson and Plotnikoff (2000)
Bennett and Cubbage (1992; Phase I)	14-d *Hyalella azteca* WST (S)	11	Negative Control ($\alpha = 0.05$)	Bennett and Cubbage (1992)
Bennett and Cubbage (1992; Phase II)	10-d *Chironomus dilutus* WST (S, G)	4	Negative Control ($\alpha = 0.05$)	Bennett and Cubbage (1992)
	14-d *Hyalella azteca* WST (S)	4	Negative Control ($\alpha = 0.05$)	Bennett and Cubbage (1992)
Landau Associates, Inc. (1993)	> 20-d *Chironomus dilutus* WST (S)	19	Negative Control ($\alpha = 0.05$)	Landau Associates, Inc. (1993)
	10-d *Hyalella azteca* WST (S)	19	Negative Control ($\alpha = 0.05$)	Landau Associates, Inc. (1993)
Cubbage (1992)	10-d *Hyalella azteca* WST (S)	9	Negative Control ($\alpha = 0.05$)	Cubbage (1992)
Tetra Tech (1993)	10-d *Hyalella azteca* WST (S)	15	Negative Control ($\alpha = 0.05$)	Unpublished analysis by MESL

Table 8. Summary of methods used to designate sediments collected in Washington State as toxic or not toxic.

Study	Toxicity Test	Number of Stations	Comparison Method Used for Designating Toxicity	Reference
Brady (1994)	10-d *Hyalella azteca* WST (S)	6	Negative Control ($\alpha = 0.05$)	Brady (1994)
ENSR Consulting (1994)	14-d *Hyalella azteca* WST (S)	4	Negative Control ($\alpha = 0.05$)	ENSR Consulting (1994)
FishPro Eng. and Env. (1991)	28-d *Hyalella azteca* WST (S)	9	Negative Control ($\alpha = 0.05$)	FishPro Eng. and Env. (1991)

d = day; WST = whole-sediment toxicity; PWT = pore-water toxicity; S = survival; G = growth; B = biomass.

Table 9. Summary of the evaluation of sediment cleanup objectives using toxicity designations in Table VIII of the Proposed SMS Rule Amendment.

Species / Endpoint / Region	Incidence of Toxicity[1]			
	7- to 14-day Toxicity Tests[2]		20- to 28-day Toxicity Tests[3]	
	No Exceedances of SCO[4]	One or More Exceedances of SCO[4]	No Exceedances of SCO[4]	One or More Exceedances of SCO[4]
Chironomus dilutus				
Survival				
UCR[5]	9 of 31 (29%)	0 of 19 (0%)	ND	ND
Washington State	ND	ND	ND	ND
Overall	**9 of 31 (29%)**	**0 of 19 (0%)**	**ND**	**ND**
Growth				
UCR[5]	2 of 31 (6.45%)	1 of 19 (5.26%)	ND	ND
Washington State	ND	ND	ND	ND
Overall	**2 of 31 (6.45%)**	**1 of 19 (5.26%)**	**ND**	**ND**
Biomass				
UCR[5]	ND	ND	ND	ND
Washington State	ND	ND	ND	ND
Overall	**ND**	**ND**	**ND**	**ND**
Hyalella azteca				
Survival				
UCR[5]	ND	ND	2 of 31 (6.45%)	4 of 19 (21.1%)
Washington State	ND	ND	ND	ND
Overall	**ND**	**ND**	**2 of 31 (6.45%)**	**4 of 19 (21.1%)**
Growth				
UCR[5]	ND	ND	2 of 31 (6.45%)	6 of 19 (31.6%)
Washington State	ND	ND	ND	ND
Overall	**ND**	**ND**	**2 of 31 (6.45%)**	**6 of 19 (31.6%)**
Biomass				
UCR[5]	ND	ND	ND	ND
Washington State	ND	ND	ND	ND
Overall	**ND**	**ND**	**ND**	**ND**

ND = no data; SCO = sediment cleanup objective; SMS = sediment management standards; UCR = Upper Columbia River.

[1] Toxicity was designated using criteria outlined in WDOE (2011).

[2] 7- to 14-d toxicity tests are defined as 10-d to 12-d tests for *Chironomus dilutus* or 7-d to 14-d tests for *Hyalella azteca* .

[3] 20- to 28-d toxicity tests are defined as 20-d tests for *Chironomus dilutus* or 28-d tests for *Hyalella azteca* .

[4] Analytes used in the analysis include metals (arsenic, cadmium, copper, lead, and zinc), total PAHs (polycyclic aromatic hydrocarbons), total PCBs (polychlorinated biphenyls), total DDTs (dichlorodiphenyltrichloroethane), bis(2-ethylhexyl) phthalate, and pentachlorophenol.

[5] Portion of the Columbia River between the international border and the Grand Coulee dam.

Table 10. Summary of the evaluation of sediment cleanup objectives.

Species / Endpoint / Region	Incidence of Toxicity[1]			
	7- to 14-day Toxicity Tests[2]		20- to 28-day Toxicity Tests[3]	
	No Exceedances of SCO[4]	One or More Exceedances of SCO[4]	No Exceedances of SCO[4]	One or More Exceedances of SCO[4]
Chironomus dilutus				
Survival				
UCR[5]	9 of 40 (22.5%)	1 of 24 (4.17%)	0 of 3 (0%)	2 of 6 (33.3%)
Washington State	0 of 1 (0%)	0 of 3 (0%)	0 of 5 (0%)	8 of 22 (36.4%)
Overall	**9 of 41 (22%)**	**1 of 27 (3.7%)**	**0 of 8 (0%)**	**10 of 28 (35.7%)**
Growth				
UCR[5]	16 of 40 (40%)	16 of 24 (66.7%)	1 of 3 (33.3%)	4 of 6 (66.7%)
Washington State	0 of 1 (0%)	0 of 3 (0%)	1 of 3 (33.3%)	2 of 5 (40%)
Overall	**16 of 41 (39%)**	**16 of 27 (59.3%)**	**2 of 6 (33.3%)**	**6 of 11 (54.5%)**
Biomass				
UCR[5]	28 of 40 (70%)	18 of 24 (75%)	ND	ND
Washington State	ND	ND	ND	ND
Overall	**28 of 40 (70%)**	**18 of 24 (75%)**	**ND**	**ND**
Hyalella azteca				
Survival				
UCR[5]	3 of 6 (50%)	11 of 28 (39.3%)	16 of 40 (40%)	14 of 24 (58.3%)
Washington State	9 of 32 (28.1%)	6 of 41 (14.6%)	3 of 4 (75%)	12 of 13 (92.3%)
Overall	**12 of 38 (31.6%)**	**17 of 69 (24.6%)**	**19 of 44 (43.2%)**	**26 of 37 (70.3%)**
Growth				
UCR[5]	ND	ND	2 of 40 (5%)	6 of 24 (25%)
Washington State	ND	ND	3 of 3 (100%)	4 of 5 (80%)
Overall	**ND**	**ND**	**5 of 43 (11.6%)**	**10 of 29 (34.5%)**
Biomass				
UCR[5]	ND	ND	5 of 40 (12.5%)	11 of 24 (45.8%)
Washington State	ND	ND	3 of 3 (100%)	5 of 5 (100%)
Overall	**ND**	**ND**	**8 of 43 (18.6%)**	**16 of 29 (55.2%)**

ND = no data; SCO = sediment cleanup objective; UCR = Upper Columbia River.

[1] Toxicity was desgnated using either the reference envelope approach or by statistical comparisons to negative control.

[2] 7- to 14-d toxicity tests are defined as 10-d to 12-d tests for *Chironomus dilutus* or 7-d to 14-d tests for *Hyalella azteca* .

[3] 20- to 28-d toxicity tests are defined as 20-d tests for *Chironomus dilutus* or 28-d tests for *Hyalella azteca* .

[4] Analytes used in the analysis include metals (arsenic, cadmium, copper, lead, and zinc), total PAHs (polycyclic aromatic hydrocarbons), total PCBs (polychlorinated biphenyls), total DDTs (dichlorodiphenyltrichloroethane), bis(2-ethylhexyl) phthalate, and pentachlorophenol.

[5] Portion of the Columbia River between the international border and the Grand Coulee dam.

Table 11. Summary of the evaluation of cleanup screening levels using toxicity designations in Table VIII of the Proposed SMS Rule Amendment.

Species / Endpoint / Region	Incidence of Toxicity[1]			
	7- to 14-day Toxicity Tests[2]		20- to 28-day Toxicity Tests[3]	
	No Exceedances of CSL[4]	One or More Exceedances of CSL[4]	No Exceedances of CSL[4]	One or More Exceedances of CSL[4]
Chironomus dilutus				
Survival				
UCR[5]	4 of 42 (9.52%)	0 of 8 (0%)	ND	ND
Washington State	ND	ND	ND	ND
Overall	**4 of 42 (9.52%)**	**0 of 8 (0%)**	**ND**	**ND**
Growth				
UCR[5]	1 of 42 (2.38%)	1 of 8 (12.5%)	ND	ND
Washington State	ND	ND	ND	ND
Overall	**1 of 42 (2.38%)**	**1 of 8 (12.5%)**	**ND**	**ND**
Biomass				
UCR[5]	ND	ND	ND	ND
Washington State	ND	ND	ND	ND
Overall	**ND**	**ND**	**ND**	**ND**
Hyalella azteca				
Survival				
UCR[5]	ND	ND	0 of 42 (0%)	0 of 8 (0%)
Washington State	ND	ND	ND	ND
Overall	**ND**	**ND**	**0 of 42 (0%)**	**0 of 8 (0%)**
Growth				
UCR[5]	ND	ND	0 of 42 (0%)	4 of 8 (50%)
Washington State	ND	ND	ND	ND
Overall	**ND**	**ND**	**0 of 42 (0%)**	**4 of 8 (50%)**
Biomass				
UCR[5]	ND	ND	ND	ND
Washington State	ND	ND	ND	ND
Overall	**ND**	**ND**	**ND**	**ND**

ND = no data; CSL = cleanup screening level; SMS = sediment management standards; UCR = Upper Columbia River.

[1] Toxicity was designated using criteria outlined in WDOE (2011).

[2] 7- to 14-d toxicity tests are defined as 10-d to 12-d tests for *Chironomus dilutus* or 7-d to 14-d tests for *Hyalella azteca* .

[3] 20- to 28-d toxicity tests are defined as 20-d tests for *Chironomus dilutus* or 28-d tests for *Hyalella azteca* .

[4] Analytes used in the analysis include metals (arsenic, cadmium, copper, lead, and zinc), total PAHs (polycyclic aromatic hydrocarbons), total PCBs (polychlorinated biphenyls), total DDTs (dichlorodiphenyltrichloroethane), bis(2-ethylhexyl) phthalate, and pentachlorophenol.

[5] Portion of the Columbia River between the international border and the Grand Coulee dam.

Table 12. Summary of the evaluation of cleanup screening levels.

Species / Endpoint / Region	Incidence of Toxicity[1]			
	7- to 14-day Toxicity Tests[2]		20- to 28-day Toxicity Tests[3]	
	No Exceedances of CSL[4]	One or More Exceedances of CSL[4]	No Exceedances of CSL[4]	One or More Exceedances of CSL[4]
Chironomus dilutus				
Survival				
UCR[5]	10 of 53 (18.9%)	0 of 11 (0%)	0 of 3 (0%)	2 of 6 (33.3%)
Washington State	0 of 4 (0%)	ND	2 of 11 (18.2%)	6 of 16 (37.5%)
Overall	**10 of 57 (17.5%)**	**0 of 11 (0%)**	**2 of 14 (14.3%)**	**8 of 22 (36.4%)**
Growth				
UCR[5]	23 of 53 (43.4%)	9 of 11 (81.8%)	1 of 3 (33.3%)	4 of 6 (66.7%)
Washington State	0 of 4 (0%)	ND	1 of 4 (25%)	2 of 4 (50%)
Overall	**23 of 57 (40.4%)**	**9 of 11 (81.8%)**	**2 of 7 (28.6%)**	**6 of 10 (60%)**
Biomass				
UCR[5]	35 of 53 (66%)	11 of 11 (100%)	ND	ND
Washington State	ND	ND	ND	ND
Overall	**35 of 53 (66%)**	**11 of 11 (100%)**	**ND**	**ND**
Hyalella azteca				
Survival				
UCR[5]	4 of 11 (36.4%)	10 of 23 (43.5%)	22 of 53 (41.5%)	8 of 11 (72.7%)
Washington State	10 of 56 (17.9%)	5 of 17 (29.4%)	4 of 5 (80%)	11 of 12 (91.7%)
Overall	**14 of 67 (20.9%)**	**15 of 40 (37.5%)**	**26 of 58 (44.8%)**	**19 of 23 (82.6%)**
Growth				
UCR[5]	ND	ND	4 of 53 (7.55%)	4 of 11 (36.4%)
Washington State	ND	ND	4 of 4 (100%)	3 of 4 (75%)
Overall	**ND**	**ND**	**8 of 57 (14%)**	**7 of 15 (46.7%)**
Biomass				
UCR[5]	ND	ND	10 of 53 (18.9%)	6 of 11 (54.5%)
Washington State	ND	ND	4 of 4 (100%)	4 of 4 (100%)
Overall	**ND**	**ND**	**14 of 57 (24.6%)**	**10 of 15 (66.7%)**

ND = no data; CSL = cleanup screening level; UCR = Upper Columbia River.

[1] Toxicity was desgnated using either the reference envelope approach or by statistical comparisons to negative control.

[2] 7- to 14-d toxicity tests are defined as 10-d to 12-d tests for *Chironomus dilutus* or 7-d to 14-d tests for *Hyalella azteca* .

[3] 20- to 28-d toxicity tests are defined as 20-d tests for *Chironomus dilutus* or 28-d tests for *Hyalella azteca* .

[4] Analytes used in the analysis include metals (arsenic, cadmium, copper, lead, and zinc), total PAHs (polycyclic aromatic hydrocarbons), total PCBs (polychlorinated biphenyls), total DDTs (dichlorodiphenyltrichloroethane), bis(2-ethylhexyl) phthalate, and pentachlorophenol.

[5] Portion of the Columbia River between the international border and the Grand Coulee dam.

Table 13. Evaluation of sediment cleanup objectives and sediment screening levels based on incidence of toxicity to *Chironomus dilutus* (endpoint: percent survival).

Substance / Region	Incidence of Toxicity based on *Chironomus dilutus* Percent Survival[1]							
	10- to 12-day Toxicity Tests				20-day Toxicity Tests			
	≤ SCO	SCO - CSL	≤ CSL	> CSL	≤ SCO	SQS - CSL	≤ CSL	> CSL
Arsenic								
UCR[2]	9 of 61 (14.8%)	1 of 3 (33.3%)	10 of 64 (15.6%)	ND	1 of 8 (12.5%)	1 of 1 (100%)	2 of 9 (22.2%)	ND
Washington	0 of 2 (0%)	0 of 2 (0%)	0 of 4 (0%)	ND	6 of 23 (26.1%)	2 of 4 (50%)	8 of 27 (29.6%)	ND
Overall	**9 of 63 (14.3%)**	**1 of 5 (20%)**	**10 of 68 (14.7%)**	**ND**	**7 of 31 (22.6%)**	**3 of 5 (60%)**	**10 of 36 (27.8%)**	**ND**
Cadmium								
UCR[2]	10 of 47 (21.3%)	0 of 15 (0%)	10 of 62 (16.1%)	0 of 2 (0%)	0 of 3 (0%)	ND	0 of 3 (0%)	2 of 6 (33.3%)
Washington	0 of 3 (0%)	0 of 1 (0%)	0 of 4 (0%)	ND	2 of 7 (28.6%)	1 of 5 (20%)	3 of 12 (25.0%)	5 of 14 (35.7%)
Overall	**10 of 50 (20%)**	**0 of 16 (0%)**	**10 of 66 (15.2%)**	**0 of 2 (0%)**	**2 of 10 (20%)**	**1 of 5 (20%)**	**3 of 15 (20%)**	**7 of 20 (35%)**
Copper								
UCR[2]	10 of 56 (17.9%)	0 of 3 (0%)	10 of 59 (16.9%)	0 of 5 (0%)	0 of 6 (0%)	0 of 1 (0%)	0 of 7 (0%)	2 of 2 (100%)
Washington	0 of 2 (0%)	0 of 2 (0%)	0 of 4 (0%)	ND	8 of 27 (29.6%)	ND	8 of 27 (29.6%)	ND
Overall	**10 of 58 (17.2%)**	**0 of 5 (0%)**	**10 of 63 (15.9%)**	**0 of 5 (0%)**	**8 of 33 (24.2%)**	**0 of 1 (0%)**	**8 of 34 (23.5%)**	**2 of 2 (100%)**
Lead								
UCR[2]	10 of 61 (16.4%)	0 of 2 (0%)	10 of 63 (15.9%)	0 of 1 (0%)	2 of 9 (22.2%)	ND	2 of 9 (22.2%)	ND
Washington	0 of 3 (0%)	ND	0 of 3 (0%)	ND	7 of 26 (26.9%)	1 of 1 (100%)	8 of 27 (29.6%)	ND
Overall	**10 of 64 (15.6%)**	**0 of 2 (0%)**	**10 of 66 (15.2%)**	**0 of 1 (0%)**	**9 of 35 (25.7%)**	**1 of 1 (100%)**	**10 of 36 (27.8%)**	**ND**
Zinc								
UCR[2]	10 of 55 (18.2%)	ND	10 of 55 (18.2%)	0 of 9 (0%)	0 of 6 (0%)	0 of 1 (0%)	0 of 7 (0%)	2 of 2 (100%)
Washington	0 of 4 (0%)	ND	0 of 4 (0%)	ND	5 of 24 (20.8%)	ND	5 of 24 (20.8%)	3 of 3 (100%)
Overall	**10 of 59 (16.9%)**	**ND**	**10 of 59 (16.9%)**	**0 of 9 (0%)**	**5 of 30 (16.7%)**	**0 of 1 (0%)**	**5 of 31 (16.1%)**	**5 of 5 (100%)**
Total PAHs[3]								
UCR[2]	10 of 56 (17.9%)	ND	10 of 56 (17.9%)	ND	ND	ND	ND	ND
Washington	ND	ND	ND	ND	7 of 25 (28%)	0 of 1 (0%)	7 of 26 (26.9%)	1 of 1 (100%)
Overall	**10 of 56 (17.9%)**	**ND**	**10 of 56 (17.9%)**	**ND**	**7 of 25 (28%)**	**0 of 1 (0%)**	**7 of 26 (26.9%)**	**1 of 1 (100%)**

Table 13. Evaluation of sediment cleanup objectives and sediment screening levels based on incidence of toxicity to *Chironomus dilutus* (endpoint: percent survival).

Substance / Region	Incidence of Toxicity based on *Chironomus dilutus* Percent Survival[1]							
	10- to 12-day Toxicity Tests				20-day Toxicity Tests			
	≤ SCO	SCO - CSL	≤ CSL	> CSL	≤ SCO	SQS - CSL	≤ CSL	> CSL
Total PCBs								
UCR[2]	10 of 56 (17.9%)	ND	10 of 56 (17.9%)	ND	ND	ND	ND	ND
Washington	ND	ND	ND	ND	0 of 5 (0%)	1 of 3 (33.3%)	1 of 8 (12.5%)	ND
Overall	**10 of 56 (17.9%)**	**ND**	**10 of 56 (17.9%)**	**ND**	**0 of 5 (0%)**	**1 of 3 (33.3%)**	**1 of 8 (12.5%)**	**ND**
Sum DDTs (*o, p′* + *p, p′*)								
UCR[2]	10 of 56 (17.9%)	ND	10 of 56 (17.9%)	ND	ND	ND	ND	ND
Washington	ND	ND	ND	ND	6 of 20 (30%)	ND	6 of 20 (30%)	ND
Overall	**10 of 56 (17.9%)**	**ND**	**10 of 56 (17.9%)**	**ND**	**6 of 20 (30%)**	**ND**	**6 of 20 (30%)**	**ND**
Bis(2-ethylhexyl) phthalate								
UCR[2]	10 of 56 (17.9%)	ND	10 of 56 (17.9%)	ND	ND	ND	ND	ND
Washington	ND	ND	ND	ND	1 of 8 (12.5%)	1 of 1 (100%)	2 of 9 (22.2%)	1 of 2 (50%)
Overall	**10 of 56 (17.9%)**	**ND**	**10 of 56 (17.9%)**	**ND**	**1 of 8 (12.5%)**	**1 of 1 (100%)**	**2 of 9 (22.2%)**	**1 of 2 (50%)**
Pentachlorophenol								
UCR[2]	10 of 56 (17.9%)	ND	10 of 56 (17.9%)	ND	ND	ND	ND	ND
Washington	ND	ND	ND	ND	1 of 8 (12.5%)	ND	1 of 8 (12.5%)	ND
Overall	**10 of 56 (17.9%)**	**ND**	**10 of 56 (17.9%)**	**ND**	**1 of 8 (12.5%)**	**ND**	**1 of 8 (12.5%)**	**ND**

ND = no data; SCO = sediment cleanup objective; CSL = cleanup screening level; PAH = polycyclic aromatic hydrocarbon; PCB = polychlorinated biphenyl; DDT = dichlorodiphenyltrichloroethane
UCR = Upper Columbia River.

[1] Toxicity was desgnated using either the reference envelope approach or by statistical comparisons to negative control.

[2] Portion of the Columbia River between the international border and the Grand Coulee dam.

[3] Total PAHs were calculated as the sum of 2-methylnaphthalene, acenaphthene, acenaphthylene, anthracene, benz(a)anthracene, benzo(a)pyrene, chrysene, dibenz(a,h)anthracene, fluoranthene, fluorene, naphthalene, phenanthrene, and pyrene.

Table 14. Evaluation of sediment cleanup objectives and sediment screening levels based on incidence of toxicity to *Chironomus dilutus* (endpoint: weight).

Substance / Region	Incidence of Toxicity based on *Chironomus dilutus* Growth (weight)[1]							
	10- to 12-day Toxicity Tests				20-day Toxicity Tests			
	≤ SCO	SCO - CSL	≤ CSL	> CSL	≤ SCO	SCO - CSL	≤ CSL	> CSL
Arsenic								
UCR[2]	29 of 61 (47.5%)	3 of 3 (100%)	32 of 64 (50%)	ND	4 of 8 (50%)	1 of 1 (100%)	5 of 9 (55.6%)	ND
Washington	0 of 2 (0%)	0 of 2 (0%)	0 of 4 (0%)	ND	2 of 7 (28.6%)	1 of 1 (100%)	3 of 8 (37.5%)	ND
Overall	**29 of 63 (46%)**	**3 of 5 (60%)**	**32 of 68 (47.1%)**	**ND**	**6 of 15 (40%)**	**2 of 2 (100%)**	**8 of 17 (47.1%)**	**ND**
Cadmium								
UCR[2]	22 of 47 (46.8%)	8 of 15 (53.3%)	30 of 62 (48.4%)	2 of 2 (100%)	1 of 3 (33.3%)	ND	1 of 3 (33.3%)	4 of 6 (66.7%)
Washington	0 of 3 (0%)	0 of 1 (0%)	0 of 4 (0%)	ND	1 of 3 (33.3%)	0 of 1 (0%)	1 of 4 (25.0%)	2 of 4 (50%)
Overall	**22 of 50 (44%)**	**8 of 16 (50%)**	**30 of 66 (45.4%)**	**2 of 2 (100%)**	**2 of 6 (33.3%)**	**0 of 1 (0%)**	**2 of 7 (28.6%)**	**6 of 10 (60%)**
Copper								
UCR[2]	26 of 56 (46.4%)	1 of 3 (33.3%)	27 of 59 (45.8%)	5 of 5 (100%)	2 of 6 (33.3%)	1 of 1 (100%)	3 of 7 (42.9%)	2 of 2 (100%)
Washington	0 of 2 (0%)	0 of 2 (0%)	0 of 4 (0%)	ND	3 of 8 (37.5%)	ND	3 of 8 (37.5%)	ND
Overall	**26 of 58 (44.8%)**	**1 of 5 (20%)**	**27 of 63 (42.9%)**	**5 of 5 (100%)**	**5 of 14 (35.7%)**	**1 of 1 (100%)**	**6 of 15 (40%)**	**2 of 2 (100%)**
Lead								
UCR[2]	29 of 61 (47.5%)	2 of 2 (100%)	31 of 63 (49.2%)	1 of 1 (100%)	5 of 9 (55.6%)	ND	5 of 9 (55.6%)	ND
Washington	0 of 3 (0%)	ND	0 of 3 (0%)	ND	2 of 7 (28.6%)	1 of 1 (100%)	3 of 8 (37.5%)	ND
Overall	**29 of 64 (45.3%)**	**2 of 2 (100%)**	**31 of 66 (47.0%)**	**1 of 1 (100%)**	**7 of 16 (43.8%)**	**1 of 1 (100%)**	**8 of 17 (47.1%)**	**ND**
Zinc								
UCR[2]	25 of 55 (45.5%)	ND	25 of 55 (45.5%)	7 of 9 (77.8%)	2 of 6 (33.3%)	1 of 1 (100%)	3 of 7 (42.9%)	2 of 2 (100%)
Washington	0 of 4 (0%)	ND	0 of 4 (0%)	ND	3 of 8 (37.5%)	ND	3 of 8 (37.5%)	ND
Overall	**25 of 59 (42.4%)**	**ND**	**25 of 59 (42.4%)**	**7 of 9 (77.8%)**	**5 of 14 (35.7%)**	**1 of 1 (100%)**	**6 of 15 (40%)**	**2 of 2 (100%)**
Total PAHs[3]								
UCR[2]	28 of 56 (50%)	ND	28 of 56 (50%)	ND	ND	ND	ND	ND
Washington	ND	ND	ND	ND	3 of 8 (37.5%)	ND	3 of 8 (37.5%)	ND
Overall	**28 of 56 (50%)**	**ND**	**28 of 56 (50%)**	**ND**	**3 of 8 (37.5%)**	**ND**	**3 of 8 (37.5%)**	**ND**

Table 14. Evaluation of sediment cleanup objectives and sediment screening levels based on incidence of toxicity to *Chironomus dilutus* (endpoint: weight).

Substance / Region	Incidence of Toxicity based on *Chironomus dilutus* Growth (weight)[1]							
	10- to 12-day Toxicity Tests				20-day Toxicity Tests			
	≤ SCO	SCO - CSL	≤ CSL	> CSL	≤ SCO	SCO - CSL	≤ CSL	> CSL
Total PCBs								
UCR[2]	28 of 56 (50%)	ND	28 of 56 (50%)	ND	ND	ND	ND	ND
Washington	ND	ND	ND	ND	1 of 5 (20%)	2 of 3 (66.7%)	3 of 8 (37.5%)	ND
Overall	**28 of 56 (50%)**	**ND**	**28 of 56 (50%)**	**ND**	**1 of 5 (20%)**	**2 of 3 (66.7%)**	**3 of 8 (37.5%)**	**ND**
Sum DDTs (*o, p′ + p, p′*)								
UCR[2]	28 of 56 (50%)	ND	28 of 56 (50%)	ND	ND	ND	ND	ND
Washington	ND	ND	ND	ND	1 of 3 (33.3%)	ND	1 of 3 (33.3%)	ND
Overall	**28 of 56 (50%)**	**ND**	**28 of 56 (50%)**	**ND**	**1 of 3 (33.3%)**	**ND**	**1 of 3 (33.3%)**	**ND**
Bis(2-ethylhexyl) phthalate								
UCR[2]	28 of 56 (50%)	ND	28 of 56 (50%)	ND	ND	ND	ND	ND
Washington	ND	ND	ND	ND	3 of 8 (37.5%)	ND	3 of 8 (37.5%)	ND
Overall	**28 of 56 (50%)**	**ND**	**28 of 56 (50%)**	**ND**	**3 of 8 (37.5%)**	**ND**	**3 of 8 (37.5%)**	**ND**
Pentachlorophenol								
UCR[2]	28 of 56 (50%)	ND	28 of 56 (50%)	ND	ND	ND	ND	ND
Washington	ND	ND	ND	ND	3 of 8 (37.5%)	ND	3 of 8 (37.5%)	ND
Overall	**28 of 56 (50%)**	**ND**	**28 of 56 (50%)**	**ND**	**3 of 8 (37.5%)**	**ND**	**3 of 8 (37.5%)**	**ND**

ND = no data; SCO = sediment cleanup objective; CSL = cleanup screening level; PAH = polycyclic aromatic hydrocarbon; PCB = polychlorinated biphenyl; DDT = dichlorodiphenyltrichloroethane
UCR = Upper Columbia River.

[1] Toxicity was desgnated using either the reference envelope approach or by statistical comparisons to negative control.

[2] Portion of the Columbia River between the international border and the Grand Coulee dam.

[3] Total PAHs were calculated as the sum of 2-methylnaphthalene, acenaphthene, acenaphthylene, anthracene, benz(a)anthracene, benzo(a)pyrene, chrysene, dibenz(a,h)anthracene, fluoranthene, fluorene, naphthalene, phenanthrene, and pyrene.

Table 15. Evaluation of sediment cleanup objectives and sediment screening levels based on incidence of toxicity to *Chironomus dilutus* (endpoint: biomass).

Substance / Region	Incidence of Toxicity based on *Chironomus dilutus* Biomass[1]							
	10- to 12-d Toxicity Tests				20-d Toxicity Tests			
	≤ SCO	SCO - CSL	≤ CSL	> CSL	≤ SCO	SCO - CSL	≤ CSL	> CSL
Arsenic								
UCR[2]	43 of 61 (70.5%)	3 of 3 (100%)	46 of 64 (71.9%)	ND	ND	ND	ND	ND
Washington	ND	ND	ND	ND	ND	ND	ND	ND
Overall	**43 of 61 (70.5%)**	**3 of 3 (100%)**	**46 of 64 (71.9%)**	**ND**	**ND**	**ND**	**ND**	**ND**
Cadmium								
UCR[2]	35 of 47 (74.5%)	9 of 15 (60%)	44 of 62 (80%)	2 of 2 (100%)	ND	ND	ND	ND
Washington	ND	ND	ND	ND	ND	ND	ND	ND
Overall	**35 of 47 (74.5%)**	**9 of 15 (60%)**	**44 of 62 (80%)**	**2 of 2 (100%)**	**ND**	**ND**	**ND**	**ND**
Copper								
UCR[2]	38 of 56 (67.9%)	3 of 3 (100%)	41 of 59 (69.5%)	5 of 5 (100%)	ND	ND	ND	ND
Washington	ND	ND	ND	ND	ND	ND	ND	ND
Overall	**38 of 56 (67.9%)**	**3 of 3 (100%)**	**41 of 59 (69.5%)**	**5 of 5 (100%)**	**ND**	**ND**	**ND**	**ND**
Lead								
UCR[2]	43 of 61 (70.5%)	2 of 2 (100%)	45 of 63 (71.4%)	1 of 1 (100%)	ND	ND	ND	ND
Washington	ND	ND	ND	ND	ND	ND	ND	ND
Overall	**43 of 61 (70.5%)**	**2 of 2 (100%)**	**45 of 63 (71.4%)**	**1 of 1 (100%)**	**ND**	**ND**	**ND**	**ND**
Zinc								
UCR[2]	37 of 55 (67.3%)	ND	37 of 55 (67.3%)	9 of 9 (100%)	ND	ND	ND	ND
Washington	ND	ND	ND	ND	ND	ND	ND	ND
Overall	**37 of 55 (67.3%)**	**ND**	**37 of 55 (67.3%)**	**9 of 9 (100%)**	**ND**	**ND**	**ND**	**ND**
Total PAHs[3]								
UCR[2]	42 of 56 (75%)	ND	42 of 56 (75%)	ND	ND	ND	ND	ND
Washington	ND	ND	ND	ND	ND	ND	ND	ND
Overall	**42 of 56 (75%)**	**ND**	**42 of 56 (75%)**	**ND**	**ND**	**ND**	**ND**	**ND**

Table 15. Evaluation of sediment cleanup objectives and sediment screening levels based on incidence of toxicity to *Chironomus dilutus* (endpoint: biomass).

Substance / Region	Incidence of Toxicity based on *Chironomus dilutus* Biomass[1]							
	10- to 12-d Toxicity Tests				20-d Toxicity Tests			
	≤ SCO	SCO - CSL	≤ CSL	> CSL	≤ SCO	SCO - CSL	≤ CSL	> CSL
Total PCBs								
UCR[2]	42 of 56 (75%)	ND	42 of 56 (75%)	ND	ND	ND	ND	ND
Washington	ND	ND	ND	ND	ND	ND	ND	ND
Overall	**42 of 56 (75%)**	**ND**	**42 of 56 (75%)**	**ND**	**ND**	**ND**	**ND**	**ND**
Sum DDTs (*o, p′ + p, p′*)								
UCR[2]	42 of 56 (75%)	ND	42 of 56 (75%)	ND	ND	ND	ND	ND
Washington	ND	ND	ND	ND	ND	ND	ND	ND
Overall	**42 of 56 (75%)**	**ND**	**42 of 56 (75%)**	**ND**	**ND**	**ND**	**ND**	**ND**
Bis(2-ethylhexyl) phthalate								
UCR[2]	42 of 56 (75%)	ND	42 of 56 (75%)	ND	ND	ND	ND	ND
Washington	ND	ND	ND	ND	ND	ND	ND	ND
Overall	**42 of 56 (75%)**	**ND**	**42 of 56 (75%)**	**ND**	**ND**	**ND**	**ND**	**ND**
Pentachlorophenol								
UCR[2]	42 of 56 (75%)	ND	42 of 56 (75%)	ND	ND	ND	ND	ND
Washington	ND	ND	ND	ND	ND	ND	ND	ND
Overall	**42 of 56 (75%)**	**ND**	**42 of 56 (75%)**	**ND**	**ND**	**ND**	**ND**	**ND**

ND = no data; SCO = sediment cleanup objective; CSL = cleanup screening level; PAH = polycyclic aromatic hydrocarbon; PCB = polychlorinated biphenyl; DDT = dichlorodiphenyltrichloroethane
UCR = Upper Columbia River.

[1] Toxicity was desgnated using either the reference envelope approach or by statistical comparisons to negative control.

[2] Portion of the Columbia River between the international border and the Grand Coulee dam.

[3] Total PAHs were calculated as the sum of 2-methylnaphthalene, acenaphthene, acenaphthylene, anthracene, benz(a)anthracene, benzo(a)pyrene, chrysene, dibenz(a,h)anthracene, fluoranthene, fluorene, naphthalene, phenanthrene, and pyrene.

Table 16. Evaluation of the sediment cleanup objectives and cleanup screening levels, based on incidence of toxicity to *Hyalella azteca* (endpoint: percent survival).

Substance / Region	Incidence of Toxicity based on *Hyalella azteca* Percent Survival[1]							
	7- to 14-day Toxicity Tests				28-day Toxicity Tests			
	≤ SCO	SQS - CSL	≤ CSL	> CSL	≤ SCO	SQS - CSL	≤ CSL	> CSL
Arsenic								
UCR[2]	9 of 17 (52.9%)	5 of 15 (33.3%)	14 of 32 (43.8%)	ND	28 of 61 (45.9%)	2 of 3 (66.7%)	30 of 64 (46.9%)	ND
Washington	10 of 46 (21.7%)	3 of 10 (30%)	13 of 56 (23.2%)	1 of 2 (50%)	6 of 8 (75%)	3 of 3 (100%)	9 of 11 (81.8%)	6 of 6 (100%)
Overall	**19 of 63 (30.2%)**	**8 of 25 (32%)**	**27 of 88 (30.7%)**	**1 of 2 (50%)**	**34 of 69 (49.3%)**	**5 of 6 (83.3%)**	**39 of 75 (52%)**	**6 of 6 (100%)**
Cadmium								
UCR[2]	7 of 10 (70%)	0 of 5 (0%)	7 of 15 (46.7%)	7 of 19 (36.8%)	22 of 47 (46.8%)	7 of 15 (46.7%)	29 of 62 (46.8%)	1 of 2 (50%)
Washington	11 of 43 (25.6%)	1 of 8 (12.5%)	12 of 51 (23.5%)	3 of 12 (25%)	3 of 4 (75%)	3 of 3 (100%)	6 of 7 (85.7%)	9 of 10 (90%)
Overall	**18 of 53 (34%)**	**1 of 13 (7.69%)**	**19 of 66 (28.9%)**	**10 of 31 (32.3%)**	**25 of 51 (49%)**	**10 of 18 (55.6%)**	**35 of 69 (50.7%)**	**10 of 12 (83.3%)**
Copper								
UCR[2]	8 of 25 (32%)	1 of 4 (25%)	9 of 29 (31%)	5 of 5 (100%)	24 of 56 (42.9%)	2 of 3 (66.7%)	26 of 59 (44.1%)	4 of 5 (80%)
Washington	14 of 60 (23.3%)	1 of 13 (7.69%)	15 of 73 (20.5%)	ND	10 of 12 (83.3%)	2 of 2 (100%)	12 of 14 (85.7%)	3 of 3 (100%)
Overall	**22 of 85 (25.9%)**	**2 of 17 (11.8%)**	**24 of 102 (23.5%)**	**5 of 5 (100%)**	**34 of 68 (50%)**	**4 of 5 (80%)**	**38 of 73 (52.1%)**	**7 of 8 (87.5%)**
Lead								
UCR[2]	14 of 29 (48.3%)	0 of 5 (0%)	14 of 34 (41.2%)	ND	27 of 61 (44.3%)	2 of 2 (100%)	29 of 63 (46%)	1 of 1 (100%)
Washington	13 of 59 (22%)	1 of 4 (25%)	14 of 63 (22.2%)	ND	9 of 11 (81.8%)	3 of 3 (100%)	12 of 14 (85.7%)	3 of 3 (100%)
Overall	**27 of 88 (30.7%)**	**1 of 9 (11.1%)**	**28 of 97 (28.9%)**	**ND**	**36 of 72 (50%)**	**5 of 5 (100%)**	**41 of 77 (53.2%)**	**4 of 4 (100%)**
Zinc								
UCR[2]	8 of 26 (30.8%)	1 of 2 (50%)	9 of 28 (32.1%)	5 of 6 (83.3%)	23 of 55 (41.8%)	ND	23 of 55 (41.8%)	7 of 9 (77.8%)
Washington	12 of 61 (19.7%)	ND	12 of 61 (19.7%)	3 of 3 (100%)	12 of 14 (85.7%)	ND	12 of 14 (85.7%)	3 of 3 (100%)
Overall	**20 of 87 (23%)**	**1 of 2 (50%)**	**21 of 89 (23.6%)**	**8 of 9 (88.9%)**	**35 of 69 (50.7%)**	**ND**	**35 of 69 (50.7%)**	**10 of 12 (83.3%)**
Total PAHs[3]								
UCR[2]	1 of 7 (14.3%)	ND	1 of 7 (14.3%)	ND	26 of 56 (46.4%)	ND	26 of 56 (46.4%)	ND
Washington	6 of 48 (12.5%)	0 of 1 (0%)	6 of 49 (12.2%)	3 of 4 (75%)	7 of 9 (77.8%)	1 of 1 (100%)	8 of 10 (80%)	6 of 6 (100%)
Overall	**7 of 55 (12.7%)**	**0 of 1 (0%)**	**7 of 56 (12.5%)**	**3 of 4 (75%)**	**33 of 65 (50.8%)**	**1 of 1 (100%)**	**34 of 66 (51.5%)**	**6 of 6 (100%)**

Table 16. Evaluation of the sediment cleanup objectives and cleanup screening levels, based on incidence of toxicity to *Hyalella azteca* (endpoint: percent survival).

Substance / Region	Incidence of Toxicity based on *Hyalella azteca* Percent Survival[1]							
	7- to 14-day Toxicity Tests				28-day Toxicity Tests			
	≤ SCO	SQS - CSL	≤ CSL	> CSL	≤ SCO	SQS - CSL	≤ CSL	> CSL
Total PCBs								
UCR[2]	ND	ND	ND	ND	26 of 56 (46.4%)	ND	26 of 56 (46.4%)	ND
Washington	0 of 15 (0%)	2 of 5 (40%)	2 of 20 (10%)	ND	4 of 5 (80%)	4 of 4 (100%)	8 of 9 (88.9%)	ND
Overall	**0 of 15 (0%)**	**2 of 5 (40%)**	**2 of 20 (10%)**	**ND**	**30 of 61 (49.2%)**	**4 of 4 (100%)**	**34 of 65 (52.3%)**	**ND**
Sum DDTs (*o, p′ + p, p′*)								
UCR[2]	ND	ND	ND	ND	26 of 56 (46.4%)	ND	26 of 56 (46.4%)	ND
Washington	3 of 34 (8.82%)	0 of 1 (0%)	3 of 35 (8.6%)	ND	3 of 3 (100%)	ND	3 of 3 (100%)	ND
Overall	**3 of 34 (8.82%)**	**0 of 1 (0%)**	**3 of 35 (8.6%)**	**ND**	**29 of 59 (49.2%)**	**ND**	**29 of 59 (49.2%)**	**ND**
Bis(2-ethylhexyl) phthalate								
UCR[2]	0 of 1 (0%)	1 of 1 (100%)	1 of 2 (50%)	ND	26 of 56 (46.4%)	ND	26 of 56 (46.4%)	ND
Washington	0 of 16 (0%)	2 of 9 (22.2%)	2 of 25 (8%)	0 of 2 (0%)	7 of 8 (87.5%)	4 of 4 (100%)	11 of 12 (91.7%)	ND
Overall	**0 of 17 (0%)**	**3 of 10 (30%)**	**3 of 27 (11.1%)**	**0 of 2 (0%)**	**33 of 64 (51.6%)**	**4 of 4 (100%)**	**37 of 68 (54.4%)**	**ND**
Pentachlorophenol								
UCR[2]	1 of 3 (33.3%)	ND	1 of 3 (33.3%)	ND	26 of 56 (46.4%)	ND	26 of 56 (46.4%)	ND
Washington	0 of 15 (0%)	ND	0 of 15 (0%)	ND	7 of 8 (87.5%)	ND	7 of 8 (87.5%)	ND
Overall	**1 of 18 (5.56%)**	**ND**	**1 of 18 (5.56%)**	**ND**	**33 of 64 (51.6%)**	**ND**	**33 of 64 (51.6%)**	**ND**

ND = no data; SCO = sediment cleanup objective; CSL = cleanup screening level; PAH = polycyclic aromatic hydrocarbon; PCB = polychlorinated biphenyl; DDT = dichlorodiphenyltrichloroethane
UCR = Upper Columbia River.

[1] Toxicity was desgnated using either the reference envelope approach or by statistical comparisons to negative control.

[2] Portion of the Columbia River between the international border and the Grand Coulee dam.

[3] Total PAHs were calculated as the sum of 2-methylnaphthalene, acenaphthene, acenaphthylene, anthracene, benz(a)anthracene, benzo(a)pyrene, chrysene, dibenz(a,h)anthracene, fluoranthene, fluorene, naphthalene, phenanthrene, and pyrene.

Table 17. Evaluation of the sediment cleanup objectives and cleanup screening levels, based on incidence of toxicity to *Hyalella azteca* (endpoint: weight).

Substance / Region	Incidence of Toxicity based on *Hyalella azteca* Growth (weight)[1]							
	7- to 14-day Toxicity Tests				28-day Toxicity Tests			
	≤ SCO	SCO - CSL	≤ CSL	> CSL	≤ SCO	SCO - CSL	≤ CSL	> CSL
Arsenic								
UCR[2]	ND	ND	ND	ND	8 of 61 (13.1%)	0 of 3 (0%)	8 of 64 (12.5%)	ND
Washington	ND	ND	ND	ND	6 of 7 (85.7%)	1 of 1 (100%)	7 of 8 (87.5%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**14 of 68 (20.6%)**	**1 of 4 (25%)**	**15 of 72 (20.8%)**	**ND**
Cadmium								
UCR[2]	ND	ND	ND	ND	6 of 47 (12.8%)	2 of 15 (13.3%)	8 of 62 (12.9%)	0 of 2 (0%)
Washington	ND	ND	ND	ND	3 of 3 (100%)	1 of 1 (100%)	4 of 4 (100%)	3 of 4 (75%)
Overall	**ND**	**ND**	**ND**	**ND**	**9 of 50 (18%)**	**3 of 16 (18.8%)**	**12 of 66 (18.2%)**	**3 of 6 (50%)**
Copper								
UCR[2]	ND	ND	ND	ND	5 of 56 (8.93%)	0 of 3 (0%)	5 of 59 (8.5%)	3 of 5 (60%)
Washington	ND	ND	ND	ND	7 of 8 (87.5%)	ND	7 of 8 (87.5%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**12 of 64 (18.8%)**	**0 of 3 (0%)**	**12 of 67 (17.9%)**	**3 of 5 (60%)**
Lead								
UCR[2]	ND	ND	ND	ND	8 of 61 (13.1%)	0 of 2 (0%)	8 of 63 (12.7%)	0 of 1 (0%)
Washington	ND	ND	ND	ND	6 of 7 (85.7%)	1 of 1 (100%)	7 of 8 (87.5%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**14 of 68 (20.6%)**	**1 of 3 (33.3%)**	**15 of 71 (21.1%)**	**0 of 1 (0%)**
Zinc								
UCR[2]	ND	ND	ND	ND	4 of 55 (7.27%)	ND	4 of 55 (7.27%)	4 of 9 (44.4%)
Washington	ND	ND	ND	ND	7 of 8 (87.5%)	ND	7 of 8 (87.5%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**11 of 63 (17.5%)**	**ND**	**11 of 63 (17.5%)**	**4 of 9 (44.4%)**
Total PAHs[3]								
UCR[2]	ND	ND	ND	ND	8 of 56 (14.3%)	ND	8 of 56 (14.3%)	ND
Washington	ND	ND	ND	ND	7 of 8 (87.5%)	ND	7 of 8 (87.5%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**15 of 64 (23.4%)**	**ND**	**15 of 64 (23.4%)**	**ND**

Table 17. Evaluation of the sediment cleanup objectives and cleanup screening levels, based on incidence of toxicity to *Hyalella azteca* (endpoint: weight).

Substance / Region	Incidence of Toxicity based on *Hyalella azteca* Growth (weight)[1]							
	7- to 14-day Toxicity Tests				28-day Toxicity Tests			
	≤ SCO	SCO - CSL	≤ CSL	> CSL	≤ SCO	SCO - CSL	≤ CSL	> CSL
Total PCBs								
UCR[2]	ND	ND	ND	ND	8 of 56 (14.3%)	ND	8 of 56 (14.3%)	ND
Washington	ND	ND	ND	ND	5 of 5 (100%)	2 of 3 (66.7%)	7 of 8 (87.5%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**13 of 61 (21.3%)**	**2 of 3 (66.7%)**	**15 of 64 (23.4%)**	**ND**
Sum DDTs (*o, p′ + p, p′*)								
UCR[2]	ND	ND	ND	ND	8 of 56 (14.3%)	ND	8 of 56 (14.3%)	ND
Washington	ND	ND	ND	ND	3 of 3 (100%)	ND	3 of 3 (100%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**11 of 59 (18.6%)**	**ND**	**11 of 59 (18.6%)**	**ND**
Bis(2-ethylhexyl) phthalate								
UCR[2]	ND	ND	ND	ND	8 of 56 (14.3%)	ND	8 of 56 (14.3%)	ND
Washington	ND	ND	ND	ND	7 of 8 (87.5%)	ND	7 of 8 (87.5%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**15 of 64 (23.4%)**	**ND**	**15 of 64 (23.4%)**	**ND**
Pentachlorophenol								
UCR[2]	ND	ND	ND	ND	8 of 56 (14.3%)	ND	8 of 56 (14.3%)	ND
Washington	ND	ND	ND	ND	7 of 8 (87.5%)	ND	7 of 8 (87.5%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**15 of 64 (23.4%)**	**ND**	**15 of 64 (23.4%)**	**ND**

ND = no data; SCO = sediment cleanup objective; CSL = cleanup screening level; PAH = polycyclic aromatic hydrocarbon; PCB = polychlorinated biphenyl; DDT = dichlorodiphenyltrichloroethane
UCR = Upper Columbia River.

[1] Toxicity was desgnated using either the reference envelope approach or by statistical comparisons to negative control.

[2] Portion of the Columbia River between the international border and the Grand Coulee dam.

[3] Total PAHs were calculated as the sum of 2-methylnaphthalene, acenaphthene, acenaphthylene, anthracene, benz(a)anthracene, benzo(a)pyrene, chrysene, dibenz(a,h)anthracene, fluoranthene, fluorene, naphthalene, phenanthrene, and pyrene.

Table 18. Evaluation of the sediment cleanup objectives and sediment screening levels, based on incidence of toxicity to *Hyalella azteca* (endpoint: biomass).

Substance / Region	Incidence of Toxicity based on *Hyalella azteca* Biomass[1]							
	7- to 14-day Toxicity Tests				28-day Toxicity Tests			
	≤ SCO	SCO - CSL	≤ CSL	> CSL	≤ SCO	SCO - CSL	≤ CSL	> CSL
Arsenic								
UCR[2]	ND	ND	ND	ND	15 of 61 (24.6%)	1 of 3 (33.3%)	16 of 64 (25%)	ND
Washington	ND	ND	ND	ND	7 of 7 (100%)	1 of 1 (100%)	8 of 8 (100%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**22 of 68 (32.4%)**	**2 of 4 (50%)**	**24 of 72 (33.3%)**	**ND**
Cadmium								
UCR[2]	ND	ND	ND	ND	10 of 47 (21.3%)	6 of 15 (40%)	16 of 62 (25.8%)	0 of 2 (0%)
Washington	ND	ND	ND	ND	3 of 3 (100%)	1 of 1 (100%)	4 of 4 (100%)	4 of 4 (100%)
Overall	**ND**	**ND**	**ND**	**ND**	**13 of 50 (26%)**	**7 of 16 (43.8%)**	**20 of 66 (30.3%)**	**4 of 6 (66.7%)**
Copper								
UCR[2]	ND	ND	ND	ND	11 of 56 (19.6%)	0 of 3 (0%)	11 of 59 (18.6%)	5 of 5 (100%)
Washington	ND	ND	ND	ND	8 of 8 (100%)	ND	8 of 8 (100%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**19 of 64 (29.7%)**	**0 of 3 (0%)**	**19 of 67 (28.4%)**	**5 of 5 (100%)**
Lead								
UCR[2]	ND	ND	ND	ND	15 of 61 (24.6%)	1 of 2 (50%)	16 of 63 (25.4%)	0 of 1 (0%)
Washington	ND	ND	ND	ND	7 of 7 (100%)	1 of 1 (100%)	8 of 8 (100%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**22 of 68 (32.4%)**	**2 of 3 (66.7%)**	**24 of 71 (33.8%)**	**0 of 1 (0%)**
Zinc								
UCR[2]	ND	ND	ND	ND	10 of 55 (18.2%)	ND	10 of 55 (18.2%)	6 of 9 (66.7%)
Washington	ND	ND	ND	ND	8 of 8 (100%)	ND	8 of 8 (100%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**18 of 63 (28.6%)**	**ND**	**18 of 63 (28.6%)**	**6 of 9 (66.7%)**
Total PAHs[3]								
UCR[2]	ND	ND	ND	ND	14 of 56 (25%)	ND	14 of 56 (25%)	ND
Washington	ND	ND	ND	ND	8 of 8 (100%)	ND	8 of 8 (100%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**22 of 64 (34.4%)**	**ND**	**22 of 64 (34.4%)**	**ND**

Table 18. Evaluation of the sediment cleanup objectives and sediment screening levels, based on incidence of toxicity to *Hyalella azteca* (endpoint: biomass).

Substance / Region	Incidence of Toxicity based on *Hyalella azteca* Biomass[1]							
	7- to 14-day Toxicity Tests				28-day Toxicity Tests			
	≤ SCO	SCO - CSL	≤ CSL	> CSL	≤ SCO	SCO - CSL	≤ CSL	> CSL
Total PCBs								
UCR[2]	ND	ND	ND	ND	14 of 56 (25%)	ND	14 of 56 (25%)	ND
Washington	ND	ND	ND	ND	5 of 5 (100%)	3 of 3 (100%)	8 of 8 (100%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**19 of 61 (31.1%)**	**3 of 3 (100%)**	**22 of 64 (34.4%)**	**ND**
Sum DDTs (*o, p′* + *p, p′*)								
UCR[2]	ND	ND	ND	ND	14 of 56 (25%)	ND	14 of 56 (25%)	ND
Washington	ND	ND	ND	ND	3 of 3 (100%)	ND	3 of 3 (100%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**17 of 59 (28.8%)**	**ND**	**17 of 59 (28.8%)**	**ND**
Bis(2-ethylhexyl) phthalate								
UCR[2]	ND	ND	ND	ND	14 of 56 (25%)	ND	14 of 56 (25%)	ND
Washington	ND	ND	ND	ND	8 of 8 (100%)	ND	8 of 8 (100%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**22 of 64 (34.4%)**	**ND**	**22 of 64 (34.4%)**	**ND**
Pentachlorophenol								
UCR[2]	ND	ND	ND	ND	14 of 56 (25%)	ND	14 of 56 (25%)	ND
Washington	ND	ND	ND	ND	8 of 8 (100%)	ND	8 of 8 (100%)	ND
Overall	**ND**	**ND**	**ND**	**ND**	**22 of 64 (34.4%)**	**ND**	**22 of 64 (34.4%)**	**ND**

ND = no data; SCO = sediment cleanup objective; CSL = cleanup screening level; PAH = polycyclic aromatic hydrocarbon; PCB = polychlorinated biphenyl; DDT = dichlorodiphenyltrichloroethane
UCR = Upper Columbia River.

[1] Toxicity was desgnated using either the reference envelope approach or by statistical comparisons to negative control.

[2] Portion of the Columbia River between the international border and the Grand Coulee dam.

[3] Total PAHs were calculated as the sum of 2-methylnaphthalene, acenaphthene, acenaphthylene, anthracene, benz(a)anthracene, benzo(a)pyrene, chrysene, dibenz(a,h)anthracene, fluoranthene, fluorene, naphthalene, phenanthrene, and pyrene.

Figures

Figure 1. Comparison of methods for classifying sediments as toxic or not toxic based on a function of control response (i.e., use of incorrect control−normalization procedure results in a larger effect required to result in designation of a sediment samples as toxic).

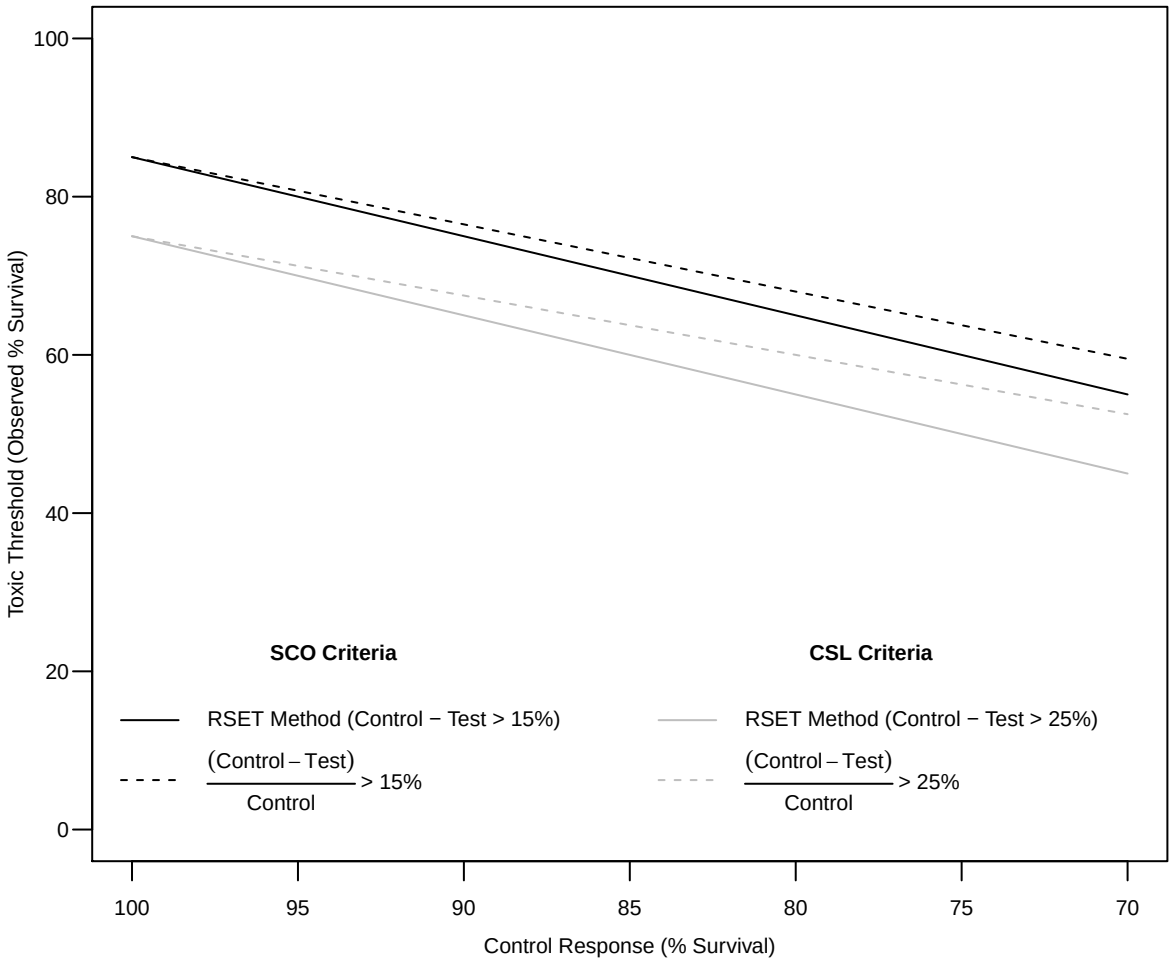


Figure 2. Relative endpoint sensitivity of *Chironomus dilutus* survival and biomass in sediments collected from the Upper Columbia River (dashed lines represent ± 20% difference from unity).

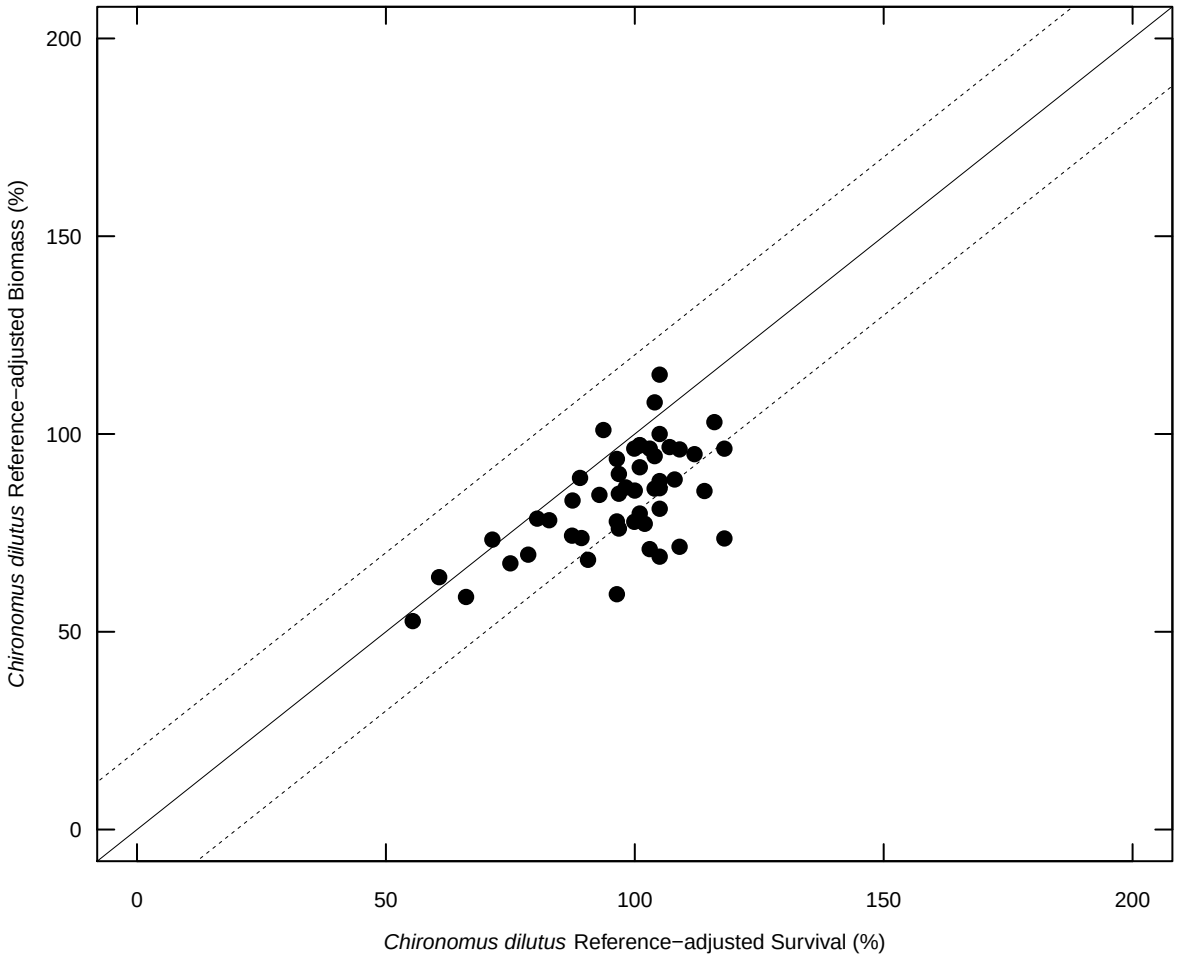


Figure 3. Relative endpoint sensitivity of *Hyalella azteca* survival and biomass in sediments collected from the Upper Columbia River (dashed lines represent ± 20% difference from unity).

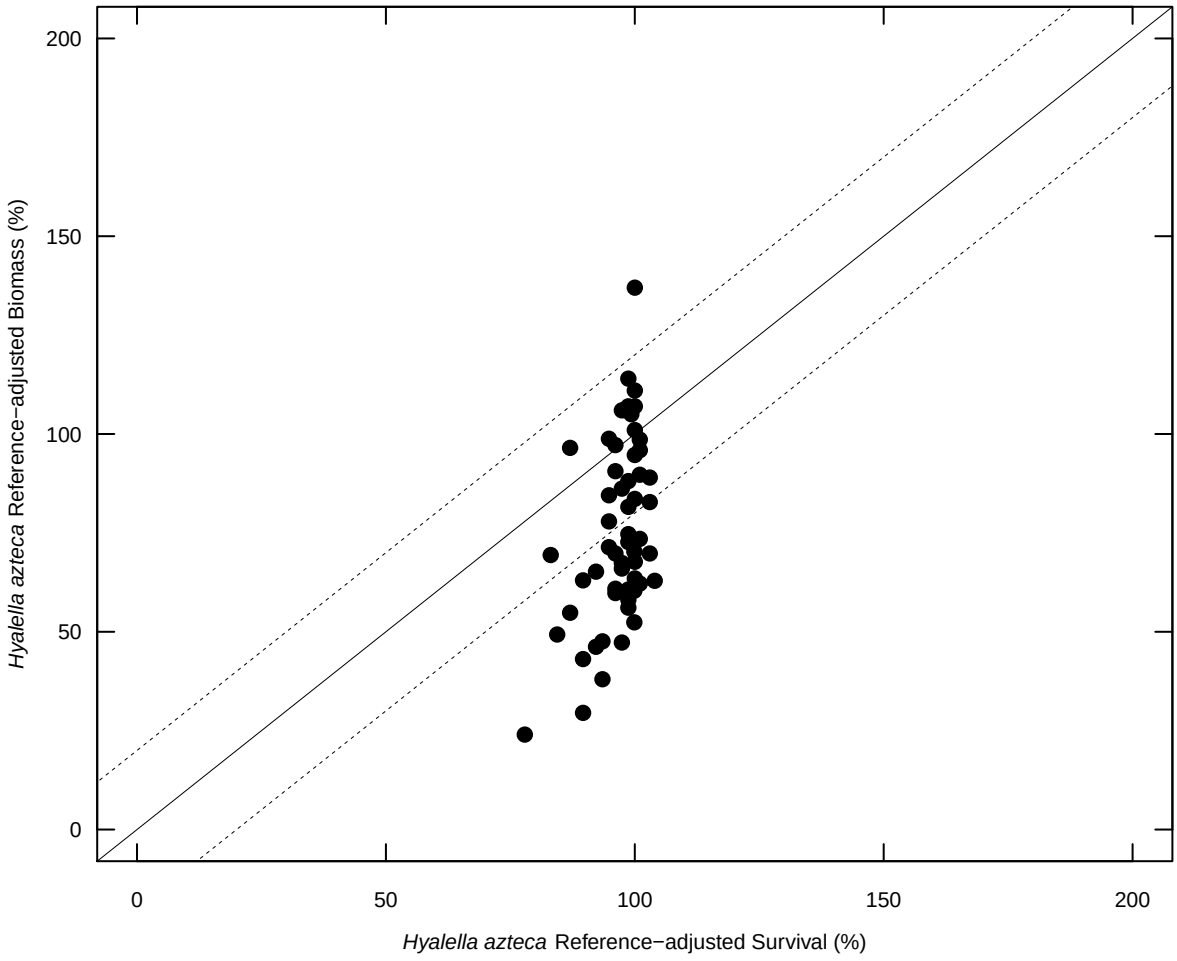
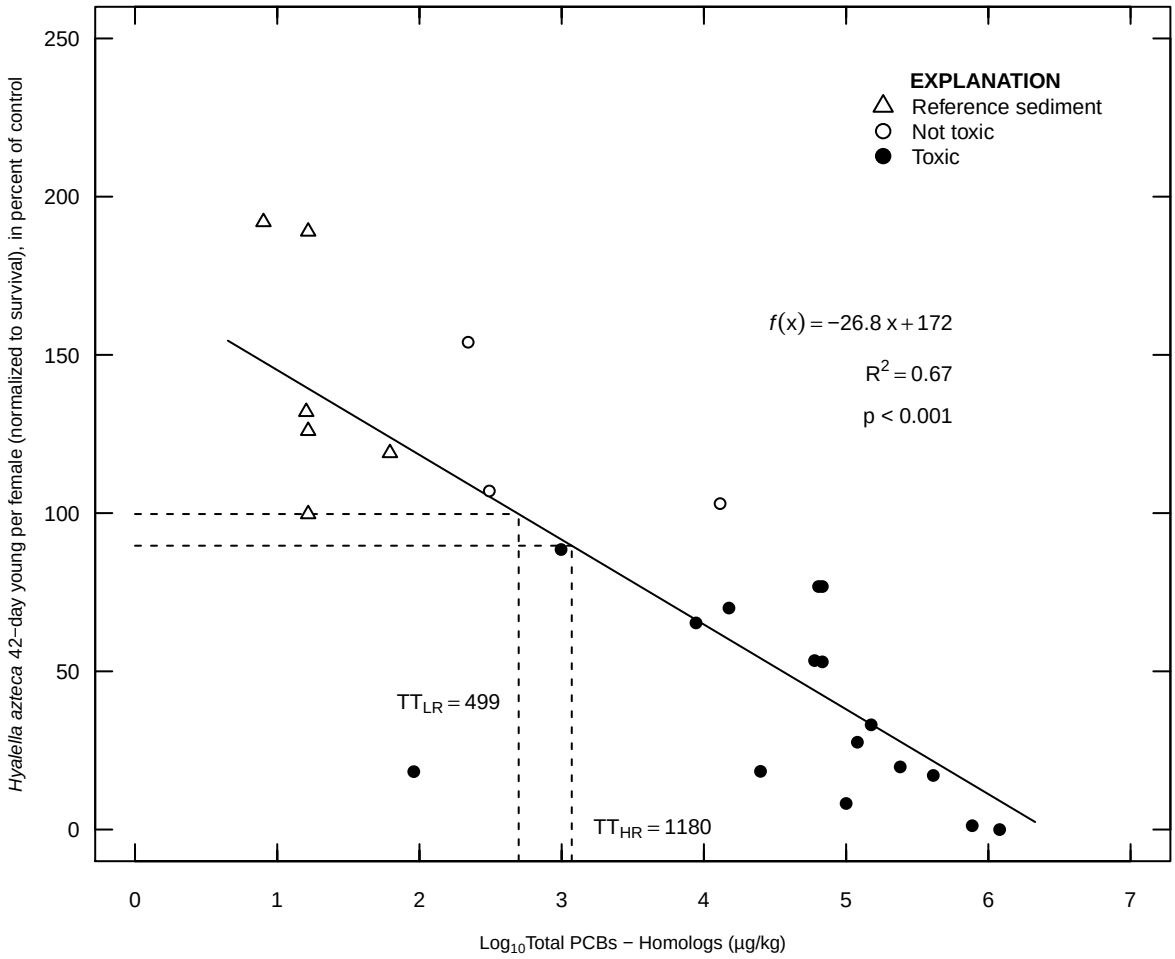


Figure 4. Concentration−response model for total polychlorinated biphenyls (PCBs; µg/kg DW) and *Hyalella azteca* reproduction (young per female normalized to percent survival) based on samples collected from the Anniston PCB site.



Appendix 1

Expert Reviewer Comments on Benthic SQVs

United States Department of the Interior

U. S. GEOLOGICAL SURVEY

Columbia Environmental Research Center
4200 New Haven Road
Columbia, Missouri 65201

Date: November 4, 2010

To: Russ McMillian

From: Chris Ingersoll (cingersoll@usgs.gov, 573-876-1819)

Subject: Review comments on the June 2010 draft report entitled "Development of Benthic SQVs for Freshwater Sediments in Washington, Oregon, and Idaho"

Listed below please find my comments on the draft report. At the end of my comments, I have briefly responded to your list of specific questions and have provided an overall summary of my comments.

Please let me know if you have any questions or if you would like to discuss my comments, particularly my brief responses to some of your specific questions.

Please note that these comments reflect my opinion, not the perspective of the USGS.

If you are interested, I would be willing to conduct an independent evaluation of the predictive ability of the draft SQVs using data sets our laboratory has compile from WA, ID, and OR (n of about 200 for 10-d midge, about 125 for 10-d amphipods, and about 75 for 28-d amphipods). I expect that I could complete such an evaluation by the end of November. Please let me know if you be interested having me provide this additional technical support.

Sincerely,

Chris Ingersoll

Dr. Chris Ingersoll
Columbia Environmental Research Center
U.S. Geological Survey
4200 New Haven Rd, Columbia, MO 65201
573/876-1819, fax -1896
cingersoll@usgs.gov
http://www.cerc.usgs.gov/

General comments:

1. Insufficient information has been provided in the report describing the process used to develop the draft WA-OR-ID sediment quality values (SQVs). Specifically, the report needs to be expanded to summarize: (1) primary toxicity or chemistry data used to develop the database and to derive the SQVs; and, (2) results of various initial analyses of the database. Without this summary information, I cannot provide a detailed review of the proposed procedure for developing regional SQVs for WA-OR-ID.

   At a minimum, the report needs to provide tables summarizing:

   a. **Number of samples included in the database by region within each of the three states.** Included in this table, there needs to be a summary of the percent of the samples that are toxic by endpoint that are: (1) significantly different from control and (2) are significantly different from control and by a reduced percentage from control (e.g., <10%, <20%, <30%).
   b. **Number of samples excluded from the database by region within each of the three states.** Included in this table, there needs to be a summary of the number of stations that were screened out of the database across all of the regions and the associated rationale for data exclusion (see Section 2.1.2).
   c. **Regional reliability of SQVs.** Section 3 of provides summaries of reliability estimates across the database by toxicity endpoint. There is a need to determine how well the SQVs estimate toxicity on a regional basis within each of the three states. See for example the summaries provided in Tables 3, 4, and 5 of Ingersoll et al. (2001; Ingersoll et al. 2001. Predictions of sediment toxicity using consensus-based freshwater sediment quality guidelines. *Arch Environ Contam Toxicol* 41:8-21).
   d. **Existing SQGs included in Table 3-3 and 3-4.** A table summarizing these existing SQGs referenced in Table 3-3 and in Table 3-4 needs to be included in the report.
   e. **Comparisons of existing SQGs to the proposed SQVs**. A table is needed comparing the relative differences in existing SQGs compared to the proposed SQVs. While some similarities are evident between some of the existing SQGs and proposed SQVs (e.g. Hg PEC of 1.06 µg/g vs. Hg SL1 of 0.66 µg/g), there are also marked differences (e.g., zinc PEC of 459 µg/g vs. Zn SL1 of 3200 µg/g and Zn SL1 of >4200 µg/g). These comparisons should be presented discussed in the report. Moreover, this table would identify chemicals of concern for which there are proposed SQVs with no existing SQGs (e.g., Tributyltin).

2. The SQVs approach has been developed using a floating percentile method (FPM). The FPM is based on the evaluating the frequency of SQV exceedances or the frequency of toxicity of samples within the database. However, the FPM as described in the report does not address the magnitude of contamination or the magnitude of toxicity. Samples are placed in one of two toxicity categories: toxic or not toxic and samples are placed in one of two chemistry categories: below an SQV or above an SQV. Over the past decade, there have been substantial advances in development and application of sediment quality guidelines (SQGs) that incorporate the magnitude of toxicity (e.g., percent survival) and the magnitude of chemistry (e.g., mean quotients) in the context assessing sediment contamination. Wenning

et al. (2005) provides examples of empirical and mechanistic SQG approaches for evaluating magnitude of toxicity or magnitude of chemical concentration. Importantly, the proposed SQV approach needs to be revised to provide mechanisms for addressing: (1) the magnitude of chemical exceedences of SQGs and (2) the magnitude of toxicity. Wenning RJ, Batley G, Ingersoll CG, Moore DW, editors. 2005. Use of sediment quality guidelines and related tools for the assessment of contaminated sediments. Pensacola FL: SETAC Press, 783 p.

3. The SQV approach based on the FPM does not adequately address the influence of mixtures or the contribution of individual chemicals of concern on observed sediment toxicity. Equal weight is given to an exceedance of a major class of compounds (e.g., total PAHs) relative to a single exceedance of a chemical (e.g., carbazole). Wenning et al. (2005) provides examples of approaches for evaluating the influence of mixtures and the influence of individual chemicals of concern associated with application of empirical or mechanistic SQGs. Importantly, the proposed SQV approach needs to be revised to provide mechanisms for more directly addressing the influence of mixtures on the predicted toxicity. The proposed SQV approach described in the report also needs to be revised to provide mechanisms for identifying the influence of individual chemicals of concern on the observed toxicity.

4. A primary objective of the report was to improve on limitations of existing SQGs (e.g., page ES1, 1st paragraph). While an evaluation of the reliability of existing SQGs was included in Tables 3-3 and 3-4, no general summary of these existing SQGs has been provided in the report, and the primary literature describing these existing SQGs has not even been cited in the report. Moreover, it is not clear as to the source of the ERLs and ERMs (marine or freshwater?). The report needs to be expanded to provide an overview of the narrative intent of existing SQGs referenced in Section 3. The report also needs to summarize these actual SQG values in an appendix to the report. Importantly, the reader needs to understand the source and narrative intent of the existing SQGs and reader needs to know how different the existing SQGs are compared to the proposed SQVs.

5. Tables 3-3 and 3-4. These two tables summarize the frequency of reliability (percentages) across different metrics. A critical piece of information missing from each of these reliability estimates is the number of samples that are used to estimate each of these frequencies. It appears from the data summarized in Table 3-2, there are a limited number of toxic samples for many of the endpoints (less than 15% across most of the endpoints?). If there are a limited number of toxic samples available for estimating false positive or false negative rates of SQVs, this limited number of toxic samples limits the utility of the reliability estimates (e.g., dramatic changes in reliability estimates can occur when there are a limited number of toxic samples within a category of interest). That said, only with sample number identified in Tables 3-3 and 3-4, can one then understand how robust particular measures of reliability might be across toxicity endpoints.

6. Tables 3-3 and 3-4. Comparisons are made between reliability of the proposed SQVs and existing SQGs. It is not clear as to how these estimates of reliability were made to the existing SQGs. Moreover, the number of toxic samples used to make each of these comparisons for each endpoint is not included in these tables (see comment above). That said, for the existing threshold-type SQGs (e.g., ERLs, TECs in Table 3-3), the frequency of false positives, are expected to be high (not intended to be predictive of toxicity). These

threshold SQGs are to have a low false negative rate which is typically the case in Table 3-3. Similarly, for the probable-type thresholds (e.g., ERMs, PECs in Table 3-4), the frequency of false negatives is expected to be high (not intended to be predictive of the lack of toxicity). What is surprising is the frequency of false positives is also relatively high for the false positives using the existing probable threshold. We have found the false positive rate to be relatively low when using probably-type SQGs to predict toxicity. Importantly, a description is needed in the report as to how the false positive and false negative rates were estimated for the existing threshold SQGs in Table 3-3 and for the exiting probable thresholds in Table 3-4 before any definitive conclusions can be made on the reported reliability of these existing SQGs in Section 3 of the report. Moreover, the report needs to cite and discuss the primary literature regarding the reliability estimates of existing SQGs compared to the reliability estimates of existing SQGs reported in Tables 3-3 and 3-4.

7. Tables 3-3 and 3-4. Overall reliability estimates the FPM values are provided in Tables 3-3 and 3-4. What is missing from the report is an evaluation of the reliability of individual SQV. It is likely that many of the individual SQVs listed in Table ES-1 would exhibit a wide range in reliability. This is a critical analysis that needs to be included in the report, particularly when individual SQVs will be used to evaluate chemicals of interest.

8. Narrative intent of SQS and SL1 values. The SL1value (and SQS) is intended to be a no observable effect level (e.g., Section 2.6.1). Given this narrative intend of an SL1 value, it is surprising that a false negative rate of 19 to 20% (Table 3-3) would meet the reliability goal of a no observable effect level. From my perspective, misclassifying 1 of every 5 samples, is not an acceptable definition of a no observable effect concentration, particularly in a database with a <20% incidence in toxicity to begin with.

9. False negative error rate. It is my understanding that a decision needs to be made before the FPM can even be run is to pre-determine an acceptable false negative rate. This decision pre-defines other measures of reliability in addition to the false negative error rate, and thus makes FPM approach more of a risk management tool rather than a risk assessment tool. Setting an allowable false negative error rate of 20%, and then allowing the FPM to increase false positive rates is not a conservative approach for protecting the environment. A maximum false negative rate of 5% or 10% would be more in line with the designed level of protectiveness of other ecological benchmarks, such as the USEPA aquatic life criteria or other types of existing SQGs.

10. Independent evaluation of predictive ability of the proposed SQVs. Comparisons between the FPM and other SQGs need to be conducted using an independent dataset. The evaluations of reliability in the report focused on evaluation using the database that was used to derive the SQVs, which may result in higher reliability of the proposed SQVs compared to using the existing SQGs to predict toxicity in the database. Since this independent predictive validation of the SQVs has not been performed in the report, the utility of the proposed SQVs compared to existing SQGs cannot be determined.

11. Page ES-1, 2nd paragraph from the bottom. The statement is made that all data was collected using ASTM-approved or Ecology-approved bioassay methods and chemistry method. A table needs to be included in the report that summarizes the key test acceptability

requirements from ASTM or from Ecology that were used to judge acceptability of the regional sediment toxicity data (see page 5, 2nd paragraph). The only acceptability requirement mentioned in the report was in reference to the number of replicates tested (5 replicates was established as the acceptability criterion in Section 2.1.2 [even though ASTM E1706 states 4 replicates is acceptable; see Section 15.2.2.4 in ASTM E1706]).

12. Page 8, 2nd paragraph. Reference is made to a "biological over-ride" of SQVs by individual toxicity tests. I do not think results of individual toxicity tests should be used in such a powerful way. Importantly, there is a need to assess sediment contamination using a "weight of evidence" approach (not trumping one robust endpoint with limited data generated from a potentially confounded single toxicity endpoint). Specifically, there should be a mechanism in the WA-OR-ID sediment assessment framework to go beyond initial screening of sediment toxicity using: (1) empirical SQVs and (2) sediment toxicity tests. For example, the use of: (1) mechanistically based SQGs (equilibrium partitioning) and (2) toxicity identification evaluations should be included in the WA-OR-ID sediment assessment program. Use of tools beyond toxicity tests and SQVs are needed when costly decisions are to be made regarding actions such as clean up of sediment or source control. See Wenning et al. (2005) for detailed discussions of the use of weight of evidence approaches for assessing potential risks of contaminants in sediment.

13. Page 8, 3rd paragraph. The statement is made that "identification of adverse biological effects generally involves a statistical difference from control or reference plus some threshold of effects." This statement is not supported by either guidance or by recommendations in either ASTM or in USEPA methods for establishing toxicity of sediments. Requiring a percentage difference from control or reference is a policy judgment that some regional groups have used to identify toxicity beyond comparisons to control sediments or reference sediments.

14. Table 2-2. Analyses are needed in the report that evaluating reliability of the SQVs relative to the various procedures used to establish toxicity relative to control conditions (e.g., significance from control alone versus significance from control and a percent difference from control).

15. Table 2-2. I do not understand the rationale for establishing separate quality assurance for response of organisms in control sediment versus reference sediment (e.g., *Hyalella azteca* 10-d testing <20% mortality required for control sediment and <25% mortality required for reference sediment). Why are these requirements not the same? What makes 25% mortality acceptable in a reference sediment?

16. Table 2-2. I do not understand why the required differences from control sediment are different from the required differences from reference sediment (e.g. *Hyalella azteca* 10-d testing 15% for control and 25% for reference). Why are these requirement not the same? What makes at a 15% difference from a control a problem compared to a 25% difference from reference a problem?

17. Table 2-2. The rationale for requiring a 20 to 40% difference in growth from control to establish toxicity in amphipods or in midge toxicity tests is not justified in the report. These

are a very large difference from control (e.g., toxic samples may be identified as non-toxic using these broad growth ranges).

18. Page 8, last paragraph to page 11, 2nd paragraph. I am completely lost by the rationale that has been used come to the following type statements for each species and each endpoint under each of the subsections on pages 8 to 11: "Given this, the maximum mortality that would observed at the SQS/SL1 level would be 30-35%...". The report needs to describe in much more detail regarding the rationale for each of these "given statements" on pages 8 to 11 for all species and endpoints.

19. Page 10, 1st paragraph. "Policy" objectives should not be used to establish recommended toxicity endpoints. See also Page 16, 2nd paragraph (error levels were set at comfort levels?).

20. Section 2.4 Exploratory Analyses and Section 2.5 Final Model Runs (page 16 to 18). The report needs to summarize the results of these exploratory and final model runs. Without these summaries, the acceptability of the decisions in Section 2.4 and in 2.5 regarding the final modeling effort cannot be reviewed.

21. Use of ANOVA to eliminate chemicals. The FPM uses an ANOVA test between hit and no hit distributions to justify the chemical lists included in and excluded from the model. The ANOVA test has an assumption of normality and may be invalid for most of the chemical distributions sediments. Most of the distributions of chemicals in sediment are highly skewed. Re-analyses are needed to determine if the chemicals eliminated exhibit statistical differences between hit and no hit concentrations using non-parametric tests.

22. Page ES-2, 2nd to last paragraph. The conclusion that the SQVs are expected to be protective of endangered species in Washington, Oregon, or Idaho has not been adequately addressed in the report. Many species of concern (e.g., lamprey in Portland Harbor, white sturgeon and mussels in the Upper Columbia River, snails in Idaho, to name a few) may be highly sensitive to contaminants of concern in sediment across the three state region. Even if there are no listed species in WA, OR, or ID that are present in areas where dredging or clean up is likely (page 18, last paragraph), the SQVs will likely be applied on a regional bases to assess sediment contamination where there may be listed species (e.g. White sturgeon inhabiting metal-contaminated sites in the Upper Columbia River).

23. Page i. A regional sediment evaluation team (RSET) provided guidance on the initial development of the SQVs. Has the draft report should be reviewed by each member of the RSET team? It is surprising that USFWS members of the RSET would agree with the conclusion that the SQVs are protective of endangered species.

24. The development of regional SQVs for WA, OR, and ID would benefit from organizing a workshop including a national sediment evaluation team to further discuss methods for (1) developing and applying both empirical and mechanistic SQGs and (2) conducting sediment toxicity assessments.

Specific comments:

25. Title: It would be more appropriate to re-title the report "**Options for establishing** benthic SQVs for freshwater sediments in WA, OR, and ID", based on the statement made on page 28, last paragraph, that this report is intending to provide agencies with options for selecting SQVs, not with absolute SQVs.

26. Page ES-1, 2nd paragraph. "Subchronic" is not a routinely used term (e.g., ASTM or in USEPA methods focus more on duration and endpoints). Moreover, calling a 28-d lethal endpoint an "acute" endpoint is not a routinely used term (e.g., Page 7, last sentence).

27. Page ES-2. The conclusions on accuracy of the proposed SQVs and the comparisons of the proposed SQVs to existing SQG has not be adequately supported by the analyses presented in Table 3-3 and in Table 3-4

28. Table ES1-1. Why is chlordane not listed?

29. Section 1.3. What national sediment experts were asked to review the SQVs developed using the FPM?

30. Section 2.1.1. Cite Appendix B in this first paragraph.

31. Section 2.1.2. Summarize the number of stations that were screened out of the database across all of the regions.

32. Page 5, 2nd paragraph. That data are "too old" is not a test acceptability criterion described in ASTM standards. Is such a criterion described in the Washington Department of Ecology standards?

33. Page 5, 3rd paragraph. Explain the rationale for requiring a minimum of 30 detected values was chosen as a requirement to include a chemical in the derivation of an SQV.

34. Section 2.1.3, 1st paragraph. The statement is made that organic carbon normalization was not done because it is difficult to understand by the regulated community. This is not a sufficient rationale for not including an evaluation of the influence of organic carbon normalization on the reliability of the proposed SQVs. If this rationale is used to exclude organic carbon normalization procedures, then one would be hard pressed to justify the use of the FPM, given how difficult the procedure is to understand.

35. Section 2.1.3, 1st paragraph. The statement is made that organic carbon analyses of the limited 2002 data did not improve reliability. These types of analyses need to be done on the expanded 2010 database, given the limitations that were identified in the 2002 database.

36. Section 2.1.3, 3rd paragraph. Summarize the analyses comparing comparisons of THP and PAHs relative to explaining petroleum toxicity, particularly for those stations where TPHs were not measured.

37. Section 2.1.4, 1st paragraph. The report has not adequately summarized the analyses leading to the conclusion that there "appears" to be no reliability advantage to estimating in the 2002 database when making toxicity designations relative to control versus relative to reference

conditions. Moreover, these analyses need to be performed using the expanded 2010 database, given the limitations identified in the report regarding the 2002 database.

38. Section 2.1.4, 1st paragraph. How would one "standardize" reference areas?

39. Table 2-2, footnote a. This footnote is applicable to all of the QA control conditions listed in Table 2-2.

40. Page 10, 4th paragraph (and throughout the report). Ingersoll et al. (2008) did not make these "recommendations', these were "suggestions".

41. Section 2.1.6. Screening of data using ANOVA. Insufficient information has been provided in the report to justify the results of these screening analyses (and Appendix B does not provide enough "detail").

42. Page 13, last paragraph. Insufficient information has been provided in the report describing how "optimized' site-specific SQVs can be developed from the SQVs summarized in Table E1-1.

43. Table 3-7. The floating percentiles in each column need to be identified relative to the specific toxicity endpoint.

44. Reference. Most all of the references are not complete (volume numbers missing, presentations but not publications cited). ASTM 2005 should be cited as ASTM 2010. USEPA (2009) is not a correct citation (it is USEPA 2003).

45. Appendix A. This table needs to be expanded (see Comment 1)

46. Appendix B. Insufficient analyses or summaries are presented in this appendix to review the data screening process (see Comment 1).

47. Page B-5. It is very surprising as to the types of chemicals that were screened out across toxicity endpoints (e.g., lots of metals of typical concern in sediment).

48. Technical memorandum (dated March 14, 2010). The results of these analyses need to be described in the report. I do not remember discussing these analyses.

49. Overview of the Biological Freshwater Sediment Standards (dated August 25, 2010): *Hyalella azteca* 10-d growth needs to be added as a sublethal endpoint. Biomass of amphipods and midge should also be added as endpoints.

Review questions posed by Russ McMillian:

(Ingersoll brief responses IN CAPITAL LETTERS; see more detailed comments above)

50. Do you agree with Ecology's conclusion that the use of sediment bioassays provides a credible basis for predicting adverse benthic effects that is consistent with current scientific information? CGI: YES.

51. Do you agree with Ecology's conclusion that multivariate statistical analysis provides a credible basis for characterizing the relationships between chemical concentrations and biological test results? CGI: YES.
52. Do you agree with Ecology's conclusion that the database provides sufficient data to support the development of statewide chemical criteria? CGI: NO. THE SOURCES OF DATA USED TO DEVELOP THE FPM ARE NOT ADEQUATELY IDENTIFIED. IMPORTANTLY, THE SQVs ARE TO BE APPLIED ACROSS REGIONS WITHIN ALL THREE STATES, NOT FOR WA STATE ALONE. INSUFFICIENT ANALYSES ARE PRESENTED IN THE REPORT TO DETERMINE IF THE PROPOSED SQVs ARE SUFFICIENTLY RELIABLE TO BE USE ON A REGIONAL BASIS WITHIN A STATE OR ACROSS THE THREE STATES.
53. That is, is the dataset sufficiently robust for the development of chemical values specific to:
    a. Geographical coverage. CGI: NO, INSUFFICIENT DATA OR INSUFFICIENT ANALYSES PRESENTED IN THE REPORT.
    b. Coverage of different types of freshwater systems. CGI: NO, INSUFFICIENT DATA OR INSUFFICIENT ANALYSES PRESENTED IN THE REPORT.
    c. Numbers of paired chemistry and bioassay endpoints. CGI: POSSIBLY, BUT IT SEEMS LIKE MOST OF THE DATA MAY BE FROM PORTLAND HARBOR, WHICH IS NOT A TYPICAL SITE BASED ON ONGOING EVALUATIONS THAT ARE BEING CONDUCTED BY USEPA ASSOCIATED WITH A BASELINE ECOLOGICAL RISK ASSESSMENT AT PORTLAND HARBOR.
    d. Number of bioassay species. CGI: YES.
    e. Number of acute and chronic tests (referring to test duration relative to life history). CGI: UNCERTAIN. MICROTOX SHOULD NOT BE INCLUDED IN THE ANALYSES. ADDITIONALLY, THE 20-D MIDGE SHOULD NOT BE ENCOURAGED UNTIL ONGOING RESEARCH TO REFINE THE 20-D MIDGE METHOD STARTING WITH 1ST INSTAR LARVAE HAS BEEN COMPLETED.
    f. Number of lethal and sublethal effects endpoints. CGI: NO. AMPHIPOD 10-D GROWTH SHOULD BE ADDED AND TOTAL BIOMASS OF AMPHIPODS AND TOTAL BIOMASS OF MIDGE SHOULD BE ADDED.
    g. Percentage of hits and no-hits for each endpoint. CGI: NO, INSUFFICIENT DATA OR INSUFFICIENT ANALYSES HAVE BEEN PRESENTED IN THE REPORT. IT APPEARS THERE IS A LOW PERCENTAGE OF TOXIC SAMPLES, WHICH COMPROMISES RELIABILITY ESTIMATES. SAMPLE NUMBER MUST BE ADDED TO ALL OF THE ESTIMATES OF RELIABILITY IN TABLES 3-3 AND 3-4.
54. Do you agree with Ecology's conclusion that data was screened using criteria acceptable for the purposes of chemical criteria development? Specifically, refer to section 2.1.2 of the technical report which includes completeness of sediment chemical analysis, minimum number of detected analytes, QA/QC of sediment chemistry and bioassay data, and elimination of chemicals that were not directly associated with toxicity. CGI: NO, INSUFFICIENT DATA OR INSUFFICIENT ANALYSES PRESENTED IN THE REPORT.
55. Do you believe that the approach used to evaluate the reliability of the criteria is consistent with current scientific methods and principles for validating criteria? CGI: NO,

INSUFFICIENT DATA OR INSUFFICIENT ANALYSES PRESENTED IN THE REPORT. RELIABILITY OF INDIVIDUAL SQVs NEEDS TO BE REPORTED.

56. What comments do you have on the completeness and relative weight that should be given to the various reliability measures used to assess the results of the model and to compare it to other SQG sets? CGI: SQVs THAT ARE INTENDED TO SERVE AS NO OBSERVED EFFECT LEVELS (SC1) SHOULD NOT HAVE A 20% FALSE NEGATIVE RATE, THIS RATE SHOULD BE BELOW ABOUT 10%. IN ADDITION, SQVs INTENDED TO REPRESENT MINOR EFFECT THRESHOLD SHOULD BE BASED ON LEVEL 1 RATHER THAN LEVEL 2 EFFECTS.
57. Are there appropriate alternate validation methods that can use the data from which the standards were developed (i.e., bootstrapping methods, etc.)? CGI: NO. INCLUDE ANALYSES CONSIDERING: (1) ORGANIC CARBON NORMALIZATION, (2) MAGNITUDE OF CHEMICAL EXCEEDANCES, (3) MAGNITUDE OF TOXICITY, AND (4) THE INFLUENCE OF DIFFERENT PERCENT DIFFERENCE FROM CONTROL ON THE PERCENT TOXIC AND ON RELIABILITY.
58. Comparative reliability analysis was used to assess different ways to handle data with respect to number of issues. For example, the toxicity of petroleum compounds can be assessed using various approaches (TPH, individual PAHs, total PAHs, normalizations, etc.), and there were other similar decisions that needed to be made regarding the underlying data set (e.g., summing, normalization, inclusion of conventionals, splitting or lumping geographic areas). Were the reliability comparisons the most appropriate method to make these decisions or are there better methods that could be used? CGI: NO, INSUFFICIENT DATA OR INSUFFICIENT ANALYSES PRESENTED IN THE REPORT.
59. Data Interpretation and Use for Regulatory Decision-Making: Greater differences in the chemical criteria occurred between the bioassays than between SQS and CSL level effects for any one bioassay endpoint. To select SQS and CSL levels for each chemical, the values for all bioassay endpoints and effects levels were combined into a single distribution representing the range of criteria from the lowest no adverse effects level to the highest minor adverse effects level. From this, the SQS was established as the lowest value and the CSL was selected as the next highest, significantly different value (see Table 3-7 of the technical report). This approach was selected as it provides conservative values by remaining at the low end of the no adverse effects to minor adverse effects distribution. Do you agree with Ecology's conclusion that this approach is consistent with that of the WA SMS marine standards where the SQS and CSL were established as the lowest and 2nd lowest of the Apparent Effects Levels determined for a suite of bioassays? CGI: I DO NOT HAVE A GOOD ENOUGH UNDERSTANDING OF THE WA SAS MARINE STANDARDS TO COMMENT ON THIS QUESTION. HOWEVER, SQVs DERIVED IN THE REPORT DO NOT REPRESENT CONSERVATIVE NO-EFFECT LEVELS OR MINOR-EFFECT LEVELS. MY EXPERIENCE IN EVALUATING DATA FROM OTHER SITES INIDICATES THAT ADVERSE EFFECTS WOULD BE COMMONLY OBSERVED AT CONENTRATIONS OF CHEMICALS OF POTENTIAL CONCERN AT OR BELOW THE SQVS.
60. Given the types of historic sediment data available, is the TPH method the best available approach for assessing the overall effects of petroleum hydrocarbons? What is your answer based on (theory or empirical data)? CGI: NO, INSUFFICIENT DATA OR INSUFFICIENT ANALYSES PRESENTED IN THE REPORT.

61. Are the measures introduced in the model to assess covariance, coupled with other available statistical tests of covariance, sufficient to address the inevitable co-occurrence of chemicals in the field when developing chemical criteria? CGI: NO ANALYSES PRESENTED IN THE REPORT REGARDING STATISTICAL ANALYSIS OF COVARIANCE. INSUFFICIENT INFORMATION PRESENTED IN THE REPORT TO EVALUATE THE INFLUENCE OF CO-OCCURRENCE IN SEDIMENT CHEMISTRY WITHIN THE DATABASE.
62. Should any chemical classes be summed that were not summed in the model to reduce covariance (e.g. phthalates)? CGI: YES. METALS SHOULD BE SUMMED AS A GROUP (SEE INGERSOLL ET AL. 2001). OTHER CHEMICAL MIXTURE MODELS SHOULD BE EVALUATED.
63. While reviewing the biological criteria, Ecology asks that reviewers consider the technical and scientific aspects of using bioassays including bioassay organisms and endpoints. Please refer to the document titled "Overview of biological freshwater sediment standards". The suite of bioassay species and endpoints were selected based partly on regional availability and familiarity with these organisms. Is the proposed bioassay suite appropriately sensitive to protect the freshwater macro benthic community (i.e., typical taxonomic structure and functions such as a prey base to endangered species like salmon)? Y CGI: YES, BUT SEE COMMENT 53E. WHOLE-SEDIMENT MUSSEL TESTING SHOULD BE CONSIDERED AS THIS METHOD BECOMES STANDARDIZED (ONGOING RESEARCH AT OUR LABORATORY).
64. From your experience, are there other freshwater bioassays/species that provide consistent, reproducible and sensitive results that should be considered for developing biological criteria? CGI: MAYFLY TESTING (ASTM E1706) AND MUSSEL TESTING (E2455).
65. Are there issues you may be familiar with regarding running and interpreting these bioassays (e.g., problems associated with culturing animals, confounding variables that may warrant protocol modifications, choice of toxicant for positive controls, etc.)? CGI: YES, AS ADDRESSED IN ASTM AND IN EPA STANDARDS. INSUFFICIENT TIME TO DISCUSS THESE TOPICS IN THE CONTEXT OF THE REPORT AND THE TIMEFRAME FOR PROVIDING REVIEW COMMENTS.
66. What comments do you have on the appropriateness of the various growth endpoints under consideration nationally (e.g., dry weight, ash-free dry weight, length)? CGI: BIOMASS OF AMPHIPODS AND BIOMASS OF MIDGE SHOULD BE ADDED AS ENDPOINTS. AFDW OF MIDGE, BUT NOT AMPHIPODS SHOULD BE REQUIRED.
67. Is there additional information on the minimum detectable difference for these tests that would assist in setting the SQS (SL1) interpretive endpoint? CGI: I DO NOT AGREE WITH THE REQUIREMENT TO EXEED A MDD IN ORDER TO IDENTIFY A SAMPLE AS TOXIC, PARTICULARLY WHEN IT IS IN THE CONTEXT OF A SL1/NO OBSERVED EFFECT LEVEL.
68. What are the pros and cons of using two endpoints from the same bioassay (e.g., the Hyalella 28 day test has two endpoints commonly reported from the same test, mortality and growth). Are there ways to maximize the use of the combined results? CGI: ADD BIOMASS AS AN ENDPOINT IN ALL MIDGE AND IN ALL AMPHIPOD TESTS.

In summary, the current draft of the subject document does not provide a scientifically defensible basis for establishing SQVs for freshwater sediments in Washington, Oregon, and Idaho. The key limitations of the approach include:

- Data sources have not been adequately described;
- Database is likely biased toward a single site (Portland Harbor, which appears to have atypical relationships between toxicity and chemistry; not representative of the tri-state region);
- A number of important toxicity test endpoints have not been considered in the analysis (e.g., biomass);
- Endpoints have not been combined to identify samples that are not toxic to any endpoint for a species (e.g., effects on 28-d survival, growth, or biomass of amphipods);
- The procedure for classifying sediment samples as toxic and not toxic is not conservative and, as such, ignores potentially important effects on benthic invertebrates. A Level 1 effect of 10-20% different from control and Level 2 effect of >20% difference from control would be more appropriate;
- SQVs for mixtures of chemicals of potential concern in sediment have not been developed. As such mixture models typically provide the most accurate bases for classifying sediment samples as toxic and not toxic, this is a major limitation of the approach;
- The procedure that was used to evaluate the reliability of the SQVs did not consider the narrative intent of the SQVs (i.e., no adverse effects and minor effects thresholds). For both types of SQVs, the goal should be to minimize the incidence of false negatives. The desired incidence of false negatives for the no effects SQVs should be <10% and for the minor effects SQVs should be <20%;
- The reliability of the individual SQVs was not evaluated;
- The conclusion that the SQVs are expected to be protective of endangered species in Washington, Oregon, or Idaho has not been adequately addressed in the report. Many species of concern (e.g., lamprey in Portland Harbor, white sturgeon and mussels in the Upper Columbia River, snails in Idaho, to name a few) may be highly sensitive to contaminants of concern in sediment across the three state region; and,
- No independent evaluation of the SQVs was conducted to determine if they can be used to accurately classify sediment samples, as toxic or not toxic, not used to derive the guidelines.

Finally, the resultant SQVs are so high that these values cannot pass "the red-face test" to represent no-effect or minor-effect thresholds. My experience at sites around the North America, including sites in WA, ID, and OR, indicates that adverse effects are likely to be observed at concentrations of chemicals of potential concern in sediment below the SQVs. In particular, use of the CSL/SL2 values as clean-up screening levels would virtually certainly result in frequently incorrectly classifying sediment samples as not toxic, when they are actually toxic to one or more benthic organisms.

If you are interested, I would be willing to conduct an independent evaluation of the predictive ability of the draft SQVs using data sets our laboratory has compile from WA, ID, and OR (n of about 200 for 10-d midge, about 125 for 10-d amphipods, and about 75 for 28-d amphipods). I

expect that I could complete such an evaluation by the end of November. Please let me know if you be interested having me provide this additional technical support.

General Review Comments on WDOE SQV Report
Dave Mount – November 30, 2010

Preface

The SQV report covers the results of a lot of work and evolution of ideas and approaches. These are summarized in broad terms in the document, but I found it very difficult to absorb all of the information presented and keep it in the forefront of my mind while considering all the elements of the approach. For those that actually lived this history, it may be hard to appreciate the difficulty someone new to the project is likely to have when hit with all of it "cold." I say this at the outset because I can see a lot of potential for making comments that suggest something was not done or considered when it in fact was, as a result of losing track of some parts of the report in attempting to synthesize things.

On a related issue, this was my first attempt to understand the inner workings of the floating point (FP) approach. Although I believe I came to an understanding of it, it is very possible there are aspects of it I don't understand, and that may lead me to make assertions about it that aren't true. I don't think that's the case, but I acknowledge it is possible.

Finally, although the supporting information contained spreadsheets with many primary data and calculations, this information was so detailed that is was extremely difficult to understand what it represented. On the other hand, the SQV report was so vague in its description of what was done, that it was difficult to evaluate the conclusions independently. It was not reasonable for me to actually re-run analyses or prepare plots independently to evaluate some of the issues I raise in my comments. It would have been helpful if there had been more in the way of plots and other intermediate analysis results to help convey the analyses that underlie the SQV development. These worksheets may seem self explanatory to those that developed them, but to a naïve reader, they are quite opaque.

General Comments

The floating point (FP) approach used for deriving empirical sediment guidelines is an interesting approach. However, I have a number of serious reservations about whether it provides the appropriate basis for sediment quality guidelines as proposed by WDOE. Further analysis may reveal that some (but likely not all) of these concerns are unfounded, but they definitely need to be evaluated before these SQVs are implemented.

For one, I found it extremely difficult to evaluate the proposed SQVs against the source data. All performance metrics are summarized in tabular form, and don't allow examination of the patterns within the data. Graphical presentations are needed to provide the reader some sense of the relationship between the SQV, and the overall distribution of chemical concentration. Some suggestions for this are provided in the comments provided below, embedded in the report, and in the tabular responses to charge questions.

There is a lot of discussion in this report that either implicitly or explicitly suggests that the FP approach identifies the chemicals responsible for sediment toxicity. I reject this assertion in the

strongest possible terms. Just like other empirical guidelines, the FP model is at the mercy of the vagaries of the underlying data, the co-occurrence of multiple chemicals typical of contaminated sediments (even where only one chemical is responsible for toxicity), and the inability to account for differing bioavailability. There is absolutely nothing in this analysis that provides evidence of causality and any such implication should be removed from the document.

Without doing TIE studies on all toxic sediments, it is not possible to know exactly how empirically-derived guidelines relate to causality. However, there are a number of reasonable inferences that can be drawn based on experiences with empirical analyses of sediment chemistry/toxicity data. For example, other SQGs have found copper concentration to be a fairly strong relationship to the frequency of sediment toxicity/non-toxicity. However, among those sediment TIE studies that have been done, there is very little (none, to my knowledge) evidence that copper is ever a cause of toxicity in typical harbor sediments. It has been found to be a cause of toxicity in sites influenced by mining or other concentrated sources of mining. This suggests that its ability to predict toxicity in mixed contamination sites is related to it being commonly associated with anthropogenic contamination of many kinds. Indeed, analyses performed by Dominic Di Toro while reviewing proposed SQGs for California showed a very high degree of co-variation among virtually all sediment contaminants. Thus, it appears likely that copper is a reasonable predictor of "dirtiness", and "dirtiness" has a good correlation to sediment toxicity, so copper is a fair correlate of the likelihood a sediment will be toxic. But the use of it as a correlate of probability of toxicity is a very different thing than using it in ways that presume that it is a cause of toxicity.

The co-variation of contaminants has serious implications for an FP approach. First, the target false negative rates for the approach are in the same range of the frequency of toxicity in the underlying data. Hence, there is not too much penalty for having a chemical-specific SQV that is well above a toxic threshold, because even if it's above the toxic threshold, the percentage of samples with false positives can't rise that high, unless the contaminant is actually <u>inversely</u> related to sediment toxicity, which is probably rare or never.

Presumably, there are contaminants in the data set, such as copper, that have a broad correlation to sediment toxicity and can predict sediment toxicity with a fair accuracy. Once one contaminant is established at an SQV that flags a substantial proportion of the toxic samples as toxic (I'm going to call this a "foundation contaminant"), there is very little pressure on any of the other contaminants to flag sediments as toxic. The only chance another contaminant (I'll call it a secondary contaminant) has of influencing the overall predictive ability is if it can predict toxicity in a sample that is low in the foundation contaminant but is still toxic. Hence, it is likely to be biased toward a high value, because the false negatives already exist (relative to the foundation chemical, so it will only improve things if it can reliably finger toxic samples (true positives) without creating false positives.

A test of this sensitivity would be to remove from consideration one of the contaminants for which the SQV is most commonly exceeded (incidentally, a table showing the frequency for which <u>individual</u> SQVs were exceeded within the toxic and non-toxic samples would help the reader understand what's driving this approach), then re-derive the values. The nature of the changes in chemical-specific values and predictive ability should be instructive as to the

underlying relationships. For example, if predictive ability of the combined model does not change much, it suggests that the removed contaminant was not as important to determining sediment toxicity as was originally suggested. If SQVs for individual chemicals change dramatically, it suggests that the SQV may not be tied to a real threshold.

The inability of SQVs that are fixed, dwt-normalized concentrations to account for varying bioavailability and/or potency is a burden that the proposed SQVs are saddled with. Particularly for PAHs, but also for metals, we know very well the extreme influence of sediment composition on bioavailability, and to propose regulatory approaches that don't address this reality just doesn't seem justifiable. PAHs present the most egregious example. Toxicity of PAHs has been unequivocally shown to be dependent on several variables, including the organic carbon content of the sediment, the degree of alkylation of the PAH mixture (where assessment is based on parent PAHs, as is done for the proposed SQVs); and the relative bioavailability of PAHs (essentially, Koc). These factors are large; sediment OC can easily range from 0.5 to 20% (40-fold), relative alkylation (ratio of alkyl+parent PAHs to parent PAHs alone) from 2 to 50 (25-fold), and Koc varies across sediments by easily a factor of 100 (and more; see Hawthorne et al. 2006 ET&C 25:2901). Since these influences are largely independent of one another, the aggregate variability in expected PAH toxicity is the product of all or 40*25*100 or about 100,000-fold.

Putting this into practical examples, calculations in our laboratory suggest that the body burden associated with a 10-d EC50 for *Hyalella* is about 25.4 umol/g (this is considerably higher than the EPA PAH ESB value of 2.24 umol/g, but is based on the specific response of *Hyalella* in 10-d exposures). If we assume an average PAH molecular weight, this equates to a value of about 7,900 ug/g OC as an expected *Hyalella* 10-d EC50. Note that this is OC-normalized and assumes that <u>all</u> PAH are included, alkyl and parent. Now let's put that in the perspective of possible field conditions. If we assume Sediment A has an OC of 0.5%, is contaminated with petroleum so the alkylation factor is high (50), and the bioavailability is not reduced by black carbon or other influences, then we expect *Hyalella* to have a 10-d EC50 of about 0.8 ug/g dwt (7900*0.005/50*1). On the other hand, sediment B is has 20% OC and is contaminated with coal tar and associate pitch, with an alkylation ratio of 2, and high Koc values that reduce bioavailability by 100-fold. In this case, we expect the 10-d EC50 for *Hyalella* to be about 80,000 ug/g (7900*0.2/2*100). That is the reality of PAH assessment. Given that, I just can't support an assessment approach that tries to capture a 100,000-fold range of toxicity using two numbers that are less than a factor of 2 apart (17 to 30 ug/g dwt).

I think it's worth pointing out that if you do this calculation with some typical values, say 1% OC, alkylation ratio of 4, and no significant bioavailability reduction, you end up with a value of about 20 ug/g dwt, which is in the range of the SL1 and SL2 values. But as is frequently the case, average only exists as a calculated value, the individual sediments are higher or lower and, in the case of PAHs, by potentially very large degrees.

With regard to organic carbon normalization, I find the argument that it is "too confusing" to be pretty hollow. First off, the WDOE marine PAH numbers are OC normalized as far as I can tell from the supporting information. It is true that some empirical SQG derivation efforts have found no substantial increase in predictive ability when PAH data are OC-normalized. But given

the backdrop of a remaining 2,500-fold variability from alkylation and bioavailability, it's not surprising that a 40-fold effect of OC normalization can't be observed amongst the remaining noise. But I don't find that a convincing reason to ignore the effects of organic carbon which have been so unequivocally demonstrated experimentally.

The same general arguments apply to many metals (e.g., Cu, Cd, Zn, etc.) whose toxicity is modulated by AVS and OC, among other factors. Multiple literature studies have clearly shown the additional clarity provided by accounting for these factors, and failing to do so by relying on dwt normalized data provides uncertainty on the order of 100-fold or more just from the typical ranges in AVS and OC. Interestingly, it is likely that if AVS and OC data were available and incorporated into a FP derivation, it is probably unlikely that predictive ability would improve over dwt analyses, because of the likelihood that these metals are indicators rather than causes of sediment toxicity in harbor-type sediments. This raises the question as to whether sediment standards should reflect the concentrations of contaminant that would actually cause sediment toxicity, or if simple empirical association, without causation, is sufficient.

Questions about what metal SQVs are responding to are further emphasized by the apparent finding (not shown in sufficient detail to evaluate) that the SQVs do not do a good job of predicting toxicity an mining sites. If they don't do a good job predicting toxicity at sites where metals are almost assuredly the cause of toxicity, doesn't that basically make the case that they are likely correlates but not causes of toxicity at non-mining sites?

The comparisons of FP results to other published SQGs is only done in the context of predictive ability. As I explain in my responses to charge questions, I don't think this comparison was done fairly. But in addition to that, I think it's important to compare the FP-based guidelines to other published guidelines directly, value to value, and this includes mechanistic guidelines (primarily EqP-based) as well as other empirical guidelines. Because there is always a risk that empirical approaches might have unintended bias as a result of quirks in their underlying data, it is valuable to compare values. If the derived SL1/2 values are greatly different than published guidelines with similar intent, I think careful consideration needs to be given as to why that is. Likewise, if the FP guidelines, which are nominally intended to reflect causation, are greatly different than what would be estimated using a mechanistic approach, then one needs to be concerned as to whether the values are at all reflective of causation, and if not, if they are consistent with the WDOEs goals for the SL1/2 values.

As referenced above, I am also concerned about the degree to which SQV derivation is influenced by the proportion of toxic samples in the database and the performance criteria. For example, the number of toxic samples in the database is about 20%; if the acceptable false negative rate is 20%, then it doesn't matter where the SQV lies, the false negative rate is likely to meet a requirement of 20% false positives. If the number of toxic samples in the database was 50%, it would exert quite a different pressure on the algorithm. This could be evaluated by redeveloping the SQVs using a database containing all of the toxic samples, but a randomly selected subset of non-toxic samples such that the total number of non-toxic samples was comparable to the number of toxic samples.

I am concerned about the degree to which the FP method can pick up effects cause by rare toxicants. Assume that a chemical was a causative toxicant, but only in a handful of samples, say 3. To aid the discussion, lets assume DDT is this toxicant. If I understand the algorithm correctly, the FP model would only pick up on DDT as an SQV-worth chemical if those three DDT-driven samples were not already classified as toxic by some other parameter such as copper. If the sediment is already a "hit" based on copper, then there is no remaining DDT "signal" for the FP approach to pick up on, because the effect of DDT is already accounted for, and it would appear that DDT is has no influence in the database.

The geographic distribution of the samples are not easily assessed from Appendix A alone, especially for someone not familiar with the locations of some of the listed studies. If there is a desire to emphasize the geographic distribution, I'd suggest a map with colored dots showing the locations of the studies and with dot diameters proportional to the number of samples included for that site.

An issue related to data coverage is the types of stressors represented by the mix of study locations. Something that would indicate the types of stressors thought to dominate at each site would help convey to the reader the nature of the data set.

The range in numbers of samples from different locations is large. This has the potential to skew the derivation of SQGs because the stressor response relationships for certain sites dominate the data (e.g., Portland harbor and Hyalella 28-d mortality). This could be evaluated by examining the specific chemicals whose SQVs trigger a prediction of toxicity and determining the pattern of chemicals across and within sites. For example, if most Portland Harbor sediments are triggered by the same chemical, then it is likely that the influence of that chemical is exaggerated by the very large number of Portland Harbor samples in the database.

One can debate terminology, but I think it is inappropriate to call 10-d sediment tests "acute". Usage is such that the term "acute" implies shorter exposures than 10 d, and usually means mortality/immobilization is the only endpoint measured. Further, it seems arbitrary to decide that a 10-d test is acute and a 20-d test is "chronic." I think it would be better to simply refer to these protocols by their durations.

This report really needs to show the SQVs in graphical form. I'd suggest "interval" plots as used by Field et al to relate the frequency distribution of toxicity compared to the SQV. This involves ordering the samples from low to high, then creating groups of samples with fixed N (e.g., 10 or 20), then plotting the percentage of toxic samples within that interval as a function of the average concentration in that interval. This helps the reader visualize the magnitude of the SQs relative to the overall range and the frequency of toxicity when the SQV is exceeded. It also helps to show which chemicals are most influential in the data sets, via the position of the SQV relative to the overall data distribution.

To help understand the relationship of the SQVs to toxicity, it would be useful to show the magnitude of response observed in each of the four boxes in the contingency table, i.e., a frequency histogram or cumulative frequency distribution for all the samples falling in that category (pred hit/hit, pred hit/no-hit, etc.)..

In terms of chemical screening, I can't determine what effect the ANOVA screen had. If a chemical was only rarely a causative toxicant, but it did cause toxicity at some sites, might it get screened out because the signal of rare toxicants is too small to be significant amongst all the other noise?

Responses to WDOE Peer Review Questions
Reviewer: Dave Mount

Section	Question	Response
1. General	Do you agree with Ecology's conclusion that the use of sediment bioassays provides a credible basis for predicting adverse benthic effects that is consistent with current scientific information?	All methods of assessing risk to benthos have the potential for either artifacts or misinterpretation, but sediment toxicity tests, used in a synthetic way with other available information, are in my opinion the best available approach.
1. General	Do you agree with Ecology's conclusion that multivariate statistical analysis provides a credible basis for characterizing the relationships between chemical concentrations and biological test results?	I believe FP can be a useful tool for gaining insight into relationships between sediment toxicity and sediment chemistry. However, I think the report reflects a misplaced belief that FP analysis indicates which sediment contaminants are "responsible" for sediment toxicity, as opposed to being correlates of toxicity which may be caused by other contaminants. Many, many statements are made in the report at either imply, or directly assert that certain contaminants are, or are not, responsible for observed toxicity. However, there is <u>nothing at all</u> in this report or the underlying data that speak to causality. Further, some other sources of information argue against such relationships. For example, even though certain metals are found by FP and many other empirical SQG approaches to be a strong correlate of sediment toxicity, TIE studies conducted in our laboratory have never indicated metals to be a cause of toxicity in sediments, with the exception of extreme contamination from very specific sources of metals, such as mining wastes. This is supported by analysis using approaches like (SEM-AVS)/foc or interstitial water analysis, which typically shows that generalized metal contamination common in harbor sediments is not at levels sufficient to attribute toxicity to those metals based on our best understanding of metal bioavailability and toxicology. This does not diminish the fact that there is a strong empirical relationship, but I think it is egregiously wrong to assert that there is a causal relationship based on empirical analysis, and/or that a regulatory approach should treat metals as the "problem" in such sediments, as opposed to (appropriately) recognizing that that level of metal contamination is often <u>associated</u> with sediment toxicity. As further evidence of this issue, the report contains the indication that maybe the proposed standards should not apply to mining sites. The basis for this was not explained, but I infer that it is because the proposed standards perform poorly for data from mining sites. Isn't this a strong indication that these SQVs are not reflecting causality, if they perform poorly at locations where it is almost certain that metals are the cause of toxicity?
2. Data Issues	Do you agree with Ecology's conclusion that the database provides sufficient data to support the development of statewide chemical criteria? That is, is the dataset sufficiently robust for the development of chemical values specific to: Geographical coverage; coverage of different types of freshwater systems; numbers of paired chemistry and bioassay endpoints; number of bioassay species; number of acute and chronic tests (referring to test duration relative to life history); number of lethal and sublethal effects endpoints; percentage of hits and no-hits for each	This is a policy determination, not a scientific one. There is no standard for sufficiency, and no matter how deep the underlying data are, there is a potential for bias and artifact to be introduced by unrecognized quirks in the underlying data. One thing I was a little surprised was not addressed in the report was the degree to which different subsets within the overall data set represented different suites of contaminants. Particularly given that different data sets have different numbers of samples, this seems potentially important. With regard to geographical coverage, I don't know of any evidence to suggest that sensitivity of sediment toxicity tests to contaminants has a geographical underpinning. The factors that influence response, such as organic carbon and AVS, can be found across great ranges within most any geographical area I'm aware of.

	endpoint.	

2. Data Issues	Do you agree with Ecology's conclusion that data was screened using criteria acceptable for the purposes of chemical criteria development? Specifically, refer to section 2.1.2 of the technical report which includes completeness of sediment chemical analysis, minimum number of detected analytes, QA/QC of sediment chemistry and bioassay data, and elimination of chemicals that were not directly associated with toxicity.	While I have no specific reason to believe it was not appropriate, the rules for screening the data were described in such general terms I don't know that it's possible to say for sure. I will say that I think it is a mistake to exclude data from sediment tests with only 4 replicates, especially without backing this decision with any quantitative analysis. Most of our laboratory tests have 4 replicates and we get detectable differences well within the bounds of the WDOE minimum effect levels. How much data is excluded because of this requirement?
3. Reliability	Do you believe that the approach used to evaluate the reliability of the criteria is consistent with current scientific methods and principles for validating criteria?	The measures that are used are relevant and have precedent in the literature. However, they represent only one form of assessment. Another important issue is the degree to which they represent true causal thresholds, since their use is predicated on an assumption that they are directly linked to the presence or absence of toxicity. Because the derivation of these numbers does not address bioavailability, there is a large, inherent error in the values, because one or two dwt based numbers simply can't account for the range in effect thresholds that exist among sediments with different characteristics relating to bioavailability. This is discussed at greater length in my general comments.
3. Reliability	What comments do you have on the completeness and relative weight that should be given to the various reliability measures used to assess the results of the model and to compare it to other SQG sets?	I am concerned about the relationship between the number of toxic samples in the database and the performance criteria. For example, the number of toxic samples in the database is about 20%; if the acceptable false negative rate is 20%, then it doesn't matter where the SQV lies, the false negative rate is likely to meet a requirement of 20% false positives. If the number of toxic samples in the database was 50%, it would exert quite a different pressure on the algorithm. This could be evaluated by redeveloping the SQVs using a database containing all of the toxic samples, but a randomly selected subset of non-toxic samples such that the total number of non-toxic samples was comparable to the number of toxic samples. The categorization of samples as simply toxic or not toxic throws away a large amount of information regarding magnitude of effect. This is important – incorrectly classifying a "skinny hit" as non-toxic is a much different error than classifying an egregiously toxic sample as non-toxic. This could be assessed by developing frequency histograms or cumulative frequency distributions for the magnitude of response observed in each of the four classifications within the predicted v observed contingency table.
3. Reliability	Are there appropriate alternate validation methods that can use the data from which the standards were developed (i.e., bootstrapping methods, etc.)?	Why not apply these SQVs to one of the large databases that extend beyond Washington State? I know of no toxicological reason why contaminant response would be regionally influenced. Alternatively, one could add the Washington data to the larger database (probably already in there) then partition the data so that there is a training data set and validation set. In that regard, the comparison of the FP SQVs to other published SQGs seems slanted to me. First, if one really believes that there is a "regional influence" to sediment toxicity, then one should redevelop the other empirical SQGs using only the Washington data to determine whether the FP approach is better. Moreover, the FP SQVs are explicitly derived explicitly to combine the various contaminants in a way complimentary to the predictive ability goals. Using other empirical SQGs in a "one hit = predicted toxic" scheme, as was apparently done, is not, in my opinion, consistent with the ways in which the developers of those other empirical SQGs would recommend evaluating the "reliability" of those SQGs for predicting sediment toxicity. My sense is that approaches like SQG "quotients" are the way those

		multiple SQG assessments are going. To be more fair, it seems to me that the "other" SQGs should be evaluated by developing an SQG quotient (or other combined metric as proposed by the developers of those methods), then using statistical methods to optimize quotients for SQV1 and SQV2 that best meet the predictive goals for those values. That would be an apples to apples comparison of the approaches.
3. Reliability	Comparative reliability analysis was used to assess different ways to handle data with respect to number of issues. For example, the toxicity of petroleum compounds can be assessed using various approaches (TPH, individual PAHs, total PAHs, normalizations, etc.), and there were other similar decisions that needed to be made regarding the underlying data set (e.g., summing, normalization, inclusion of conventionals, splitting or lumping geographic areas). Were the reliability comparisons the most appropriate method to make these decisions or are there better methods that could be used?	There are many issues here. Summing PAH concentrations makes toxicological sense, but the document did not describe what would be a minimum set of PAHs that has to be measured for a "total PAH" number to be defined, nor what would be done if a larger set of PAHs (e.g., alkyl PAHs) were measured. Further, it's not clear that there's any mechanism for distinguishing between PAH sites that have high or low degrees of alkylation, but it is definitively known that this affects the potency of PAH mixtures. Organic carbon normalization was discarded, which flies in the face of a large body of literature, and I don't believe the effect of organic carbon normalization was ever examined with the current data set. Even if it doesn't improve predictive ability, that is not an indication that it "doesn't matter", what it means is that there are enough other factors that influence response (e.g., black carbon and other influences on bioavailability) that prevent the organic carbon effect from being recognized by the analysis. Hence, I think it should still be done. I find the "it's too complicated for regulators" argument to be pretty thin. I absolutely would not separate geographical areas; if the SQV values were regionally dependent, it would be an indication of an unrecognized bias in the underlying data, since there is nothing "regional" about toxicology.
4. Interpretation	Do you agree with Ecology's conclusion that this approach is consistent with that of the WA SMS marine standards where the SQS and CSL were established as the lowest and 2nd lowest of the Apparent Effects Levels determined for a suite of bioassays?	I have no opinion about this as it would take considerable additional analysis to make such an evaluation.
4. Interpretation	Given the types of historic sediment data available, is the TPH method the best available approach for assessing the overall effects of petroleum hydrocarbons? What is your answer based on (theory or empirical data)?	I'm not clear on what is meant by the "TPH method." We have demonstrated experimentally that aliphatic hydrocarbons can induce toxicity to Hyalella and Chironomus separate from the effects of PAHs, so it is very appropriate to include some expression of non-PAH petroleum hydrocarbons among the constituents for which SQV are developed. However, it is not yet clear what the best measure or expression of TPH is with regard to predicting toxicity. I am a little concerned about how the FPM would treat TPH when diesel range and residual are separate variables, as the SQV for one might be raised when it is compensated for by the other. Was there a model run when the two were summed? Our experimental data indicate that biomass of both Hyalella and Chironomus are reduced by about 40% at 500 mg/kg aliphatic hydrocarbons (in the form of mineral oil). Based on that, the SQV1 of 340 mg/kg TPH-diesel seems close to right, maybe a little high, while the TPH-residual of more than an order of magnitude higher seems too high, though as I said we don't yet know exactly how best to define the toxicological potency of aliphatic hydrocarbons across a range of molecular weights.
4. Interpretation	Are the measures introduced in the model to assess covariance, coupled with other available statistical tests of covariance, sufficient to address the inevitable co-occurrence of chemicals in the field when developing chemical criteria?	Virtually all contaminants show some level of co-occurrence, and I am concerned that the algorithm used may allow SQVs for some contaminants to become artificially high (i.e., exceed a true causal threshold) because other contaminants "cover" for them. This is of particular concern for contaminants whose toxic threshold can be expected to vary widely as a function of bioavailability, such as PAHs and metals. As an example, there are many examples of sediments with PAH-induced toxicity at concentrations well below the PAH SQVs derived here. Yet there are also many examples of sediments with low toxicity of PAHs at very high concentrations because of low bioavailability. Under the FP method, it is possible for

		the PAH SQV to be allowed to be higher than a true causal threshold because the model wants to eliminate false positives, and the co-occurrence of other chemicals in the "lower but still toxic" PAH samples allows SQVs for other parameters to "cover" the occurrence of toxicity at lower concentrations. This is why it is so important to cross check these empirical SQVs against the known toxicity of these same chemicals to determine how well they agree, and when they don't, to decide what is causing the discrepancy and how best to address it.
4. Interpretation	Should any chemical classes be summed that were not summed in the model to reduce covariance (e.g. phthalates)?	No strong opinions.
Biological Criteria	Is the proposed bioassay suite appropriately sensitive to protect the freshwater macro benthic community (i.e., typical taxonomic structure and functions such as a prey base to endangered species like salmon)?	I don't know that there is a great deal of hard evidence that can be brought to bear on this question, but I think it is an accepted presumption that this is true.
Biological Criteria	From your experience, are there other freshwater bioassays/species that provide consistent, reproducible and sensitive results that should be considered for developing biological criteria?	None that have the depth of experience and interpretation behind them. An exception is that I don't see why weight is not included as an endpoint for 10-d *Hyalella* tests (or better yet, combined with survival to calculate a 10-d biomass endpoint). This is not to say that additional tests might not be added in the future as additional work is completed.
Biological Criteria	Are there issues you may be familiar with regarding running and interpreting these bioassays (e.g., problems associated with culturing animals, confounding variables that may warrant protocol modifications, choice of toxicant for positive controls, etc.)?	There is ongoing work in regard to the appropriate foods and feeding rate for 28-d *Hyalella* tests; there is reason to believe that the standard feeding rate is insufficient to allow maximum growth (Mount et al. unpublished data; some of these data were presented by Ingersoll et al. at the 2010 SETAC meeting). Although we don't yet have a definitive alternative to recommend, we believe we are close to having one and will circulate that as soon as we have it. I'm not aware of a positive control for sediment testing that has any wide use or acceptance.
Biological Criteria	What comments do you have on the appropriateness of the various growth endpoints under consideration nationally (e.g., dry weight, ash-free dry weight, length)?	Rather than assess growth and mortality separately for a given sediment toxicity test, I would strongly recommend using a biomass approach (survival*weight of survivors) to reduce each toxicity test to a single endpoint. The statements made in the report about length being more sensitive than weight are absolutely incorrect. This incorrect assertion is based on a smaller CV for the length measurement. However, the issue is the absolute SD relative to the magnitude of the difference between control and test treatments, and this is small for length. Weight increases with the cube of length, so even though the CV for weight is larger, the dynamic range of the measure is also much larger, such that the power is not different. I can expand on this argument with examples (and a re-analysis of data to show that Steevens et al. 1993 were incorrect in their assertions about length) if needed. No preference should be expressed for length over weight. Weight is also much less time-intensive (and therefore cheaper) to measure than length. <u>Only</u> AFDW should be accepted for *Chironomus*. Dry weight is subject to considerable bias from gut contents. Dwt or AFDW is acceptable for Hyalella. I'm not aware of any evidence to suggest that the choice makes a difference in terms of detecting differences in weight, although given that the dwts are already low, doing AFDW may just introduce additional noise to the measurement.

Biological Criteria	Is there additional information on the minimum detectable difference for these tests that would assist in setting the SQS (SL1) interpretive endpoint?	For both SL1 and SL2, I think there should be a maximum level of effect that is counted as indicating toxicity regardless of statistical significance. I'm OK with it being both sig diff and a minimum level at the low end, but when effects are large enough, I think sig diff becomes irrelevant. For Hyalella, the SL1 criterion for survival is larger (requires more effect) for a 10-d test than for a 28-d test. This makes no sense. I was very surprised that there was no analysis of minimum significant differences within the WDOE database to show what degree of difference was typically needed for those data.
Biological Criteria	What are the pros and cons of using two endpoints from the same bioassay (e.g., the Hyalella 28 day test has two endpoints commonly reported from the same test, mortality and growth). Are there ways to maximize the use of the combined results?	There is no downside to using both survival and growth data from the same assay, but I would strongly recommend converting to a biomass endpoint which encompasses both effects. This would simultaneously eliminate any awkwardness of having multiple endpoints from a single test.

General Comments

The SQV report represents a significant level of effort in assembling a high quality database with matching chemistry and toxicity for multiple test endpoints. The floating point models derived from this database are based on a systematic approach and are designed to provide threshold values for management based on several measures of reliability. The comments below provide some recommendations for improving the report.

Database Description

The report should include better description of the chemical and toxicity data used in model development both overall and by endpoint. I recommend adding tables showing chemical distributions by endpoint, including concentration ranges/percentiles of detected data, the maximum non-toxic concentration, and the number of toxic samples.

The report did not include sufficient detailed information to evaluate the geographic coverage or the extent to which database represented different types of freshwater systems. Although the geographic coverage or the type of freshwater system should not have much impact on the underlying concentration-response relationships, the chemical gradients represented by the different study areas are likely to be important in defining the threshold values. For example, data from the Portland Harbor Superfund site appears to represent a significant part of the database for several endpoints. The extent to which the Portland Harbor data influence the threshold values or are different from the rest of the data in the database should be evaluated.

Sublethal Endpoints

Two sublethal endpoints, *Chironomus* 10-day growth and *Hyalella* 28-day growth, were included in the FPM development, although the database for *Hyalella* was much smaller due to the exclusion of data from Portland Harbor. As a result of the decision to exclude the Portland Harbor data for *Hyalella* growth, the PAH FPM values for *Hyalella* growth were ignored in developing final SQVs because they were "unduly influenced by having no TPH in their models." The basis for excluding the Portland Harbor data for *Hyalella* growth should be presented in detail (agreement within the Work Group or with LWG is not an explanation), since EPA used the data for that endpoint for the Portland Harbor Superfund evaluation. The minimal database for *Hyalella* growth results and the fact that the FPM values for a major class of chemicals (PAHs) for that endpoint was ignored means that the midge growth is the only sublethal endpoint in the current set of freshwater benthic SQVs.

I recommend not using the growth endpoint as a standalone measure of the sublethal effects. In my view, using a "growth or survival" endpoint or a biomass endpoint is a better measure of sublethal effects. A test sample with significantly reduced survival has reduced power in detecting growth differences in the surviving organisms. In addition, the growth of survivors is not independent of the number of organisms surviving.

Model Performance Evaluation

Model performance was evaluated exclusively by reliability metrics. More information should be presented to give a more complete picture of the models, including the reliability of individual chemical models, the impact of the low prevalence of toxicity on the reliability metrics for all endpoints, and overall performance relative to the magnitude of toxicity.

The reliability of individual chemicals for the selected values should be reported, including the number of non-toxic and toxic samples above the selected threshold. Although applied as a set of chemical values, individual chemical models will determine the classification of samples that exceed the guidelines. It would also be useful to tabulate the number of samples by chemical and endpoint where an individual chemical model is the only chemical exceeding its threshold.

Low prevalence (less than 20%) of hits in the database for all endpoints has significant implications for the reliability evaluation and the development of reliable SQVs. As a result, small changes can have large impact on the reliability metrics used to determine model performance. Other metrics (eg., positive likelihood, kappa) that are either independent of the prevalence of hits or take it into account should be included in the evaluation. One way to assess the importance of this highly unequal distribution would be to randomly divide the non-toxic samples into 3 or 4 groups and pair with all of the toxic samples for SQV and reliability comparisons. Large differences in the set of FPM values or the values themselves differ significantly would be a cause for concern.

Evaluation of the overall model performance relative to the magnitude of toxicity would be helpful. One way this could be accomplished would be to plot the magnitude of toxicity by endpoint against the FPM values expressed as a maximum ratio of the chemical concentration to the FPM for individual samples.

Application of the Benthic SQVs

Any application of the benthic SQVs without validation with independent data should be conservative to avoid the potential for unusual conditions associated with non-toxicity for a small number of samples that result in FPM values at the upper end of the concentration range.

The description of "No Observable Effects" levels and "Minor Adverse Effects" levels potentially allow relatively high levels of toxicity to be classified as "no-hits". Because of the potential for extreme values to be classified as no-hits, I would not recommend using these values in ecological risk assessment.

Other Issues

The report should be edited to remove the implication that FPM threshold concentrations represent causal values.

Undocumented assertions about the views of other agencies should either be removed or the basis more fully explained and documented.

General Approach	
Do you agree with Ecology's conclusion that the use of sediment bioassays provides a credible basis for predicting adverse benthic effects that is consistent with current scientific information?	Yes, with limitations. Sediment bioassays currently represent the best available approach for evaluating adverse effects on benthic macroinvertebrates. Their credibility would be enhanced with the development and standardization of other non-lethal endpoints and other test species. Relying on sediment bioassays for predicting adverse benthic effects requires a high quality database of matching chemistry and toxicity and high standards for test quality to minimize false negatives and false positives.
Do you agree with Ecology's conclusion that multivariate statistical analysis provides a credible basis for characterizing the relationships between chemical concentrations and biological test results?	Yes, empirical models have a long track record in characterizing concentration-response relationships. However, the strength and reliability of these models is increased with evaluation of data independent from the development database.

Data Issues	
Do you agree with Ecology's conclusion that the database provides sufficient data to support the development of statewide chemical criteria? That is, is the dataset sufficiently robust for the development of chemical values specific to: • Geographical coverage. • Coverage of different types of freshwater systems. • Numbers of paired chemistry and bioassay	The report did not provide a detailed breakdown by endpoint for the geographic areas or the different types of systems, although I would not expect different concentration-response relationships in different geographical areas or different types of systems. However, it appears that a large percentage of the data for *Hyalella* 28d survival and *Chironomus* survival/growth came from the Portland Harbor Superfund site. Assuming the chemical gradients in Portland Harbor are consistent with other areas, this may not be an issue. It could be a problem if most of the data for specific chemicals came from one area (TPH?). For

endpoints. • Number of bioassay species. • Number of acute and chronic tests (referring to test duration relative to life history). • Number of lethal and sublethal effects endpoints. • Percentage of hits and no-hits for each endpoint.	Chironomus, sufficient data are available to compare chemical distributions from Portland Harbor with the distributions in the remainder of the database. A table showing the range and number of toxicity values that were classified as non-toxic and toxic for each endpoint and screening level would be useful. The database for *Hyalella* 28d growth is very limited and probably not sufficient to develop models that can be applied with confidence to new data. The rationale for excluding the data for *Hyalella* growth/biomass from Portland Harbor was not clearly explained, especially considering EPA used these data in the Superfund ecological risk assessment and feasibility study. The lack of TPH data in the *Hyalella* 28d growth database was used as a reason to ignore the PAH models for that endpoint in the selection of the final SL1 and SL2 values. Given the lack of *Hyalella* 10d growth and limited data on 28d growth, *Chironomus* 10d growth was the only non-lethal endpoint with sufficient data. The growth endpoint was treated as independent from the survival endpoint, although that is not likely to be the case. The biomass endpoint, which takes survival into account, may be a more useful way to assess sublethal effects. The development of chemical thresholds is limited by the low percentage of hits for all endpoints. The impact of the low prevalence of toxicity on the reliability measures was not discussed in the report.
Do you agree with Ecology's conclusion that data was screened using criteria acceptable for the	In general, data screening was appropriate and apparently systematic.

purposes of chemical criteria development? Specifically, refer to section 2.1.2 of the technical report which includes completeness of sediment chemical analysis, minimum number of detected analytes, QA/QC of sediment chemistry and bioassay data, and elimination of chemicals that were not directly associated with toxicity.	Other than vague references to an "insufficient analyte list", the criteria for completeness of the chemical analysis were not clearly explained. Similarly, the rules for calculating sums were not explained in the document, particularly whether a specific number of PAHs were required to calculate a sum. For the primary bioassay endpoints, ASTM considers 4 replicates acceptable, so it is not clear why they were excluded. If variance is an issue, that could be addressed. Also, it was not clear that non-toxic samples were evaluated for sufficient power to detect a significant difference. Because of the importance that individual non-toxic samples may have in the FPM approach, such an evaluation should be considered.

Reliability Testing of the Chemical Criteria	
Do you believe that the approach used to evaluate the reliability of the criteria is consistent with current scientific methods and principles for validating criteria?	The reliability metrics used are commonly used values, but have serious limitations given the low prevalence of toxicity for all endpoints and screening levels. Other approaches should be included. According to Shepard (2010), endpoints with low prevalence of toxicity (as observed in the freshwater database used) can be "made to meet reliability goals (overall reliability, predicted no-hit reliability and false positive rate) merely by raising sediment quality benchmarks so high that most or all toxic stations are incorrectly classified as nontoxic." Shepard recommends the use of other metrics such as the Kappa statistic that either take the prevalence of toxicity into account or are independent of prevalence.

	The reliability measures for individual chemicals should be presented for the selected models and screening levels. Although the individual chemical models are applied as a group, there are likely to be cases where a single chemical model is driving the classification. It would also be useful to show the number of cases for each chemical where that chemical model alone was responsible for the sample classification.
What comments do you have on the completeness and relative weight that should be given to the various reliability measures used to assess the results of the model and to compare it to other SQG sets?	The completeness and relative weight of reliability measures depend on the narrative intent of the SQG set. If the goal is to be protective of benthic effects, then a low false negative rate and high predicted no-hit reliability may be more important than false positive rate or predicted hit reliability.
Are there appropriate alternate validation methods that can use the data from which the standards were developed (i.e., bootstrapping methods, etc.)?	The best validation approach would be to apply the values to an independent data set. Without an independent data set, it would be possible to evaluate the impact of the low prevalence of toxicity on the reliability measures by adjusting the relative proportion of toxic and non-toxic samples in the database and deriving new models. For example, assuming 80/20 non-toxic to toxic distribution, it would be possible to randomly divide the non-toxic samples into 4 batches to combine with the toxic samples for comparative model development and reliability.

Comparative reliability analysis was used to assess different ways to handle data with respect to number of issues. For example, the toxicity of petroleum compounds can be assessed using various approaches (TPH, individual PAHs, total PAHs, normalizations, etc.), and there were other similar decisions that needed to be made regarding the underlying data set (e.g., summing, normalization, inclusion of conventionals, splitting or lumping geographic areas). Were the reliability comparisons the most appropriate method to make these decisions or are there better methods that could be used?	Reliability comparisons are useful but they only go so far. The underlying datasets for the different approaches differ, so the comparisons are not straightforward. The reliability comparisons were not presented directly, so it was not clear how these were carried out. It would make sense to focus the comparisons on predicted hit reliability and percent of false positives, rather than overall reliability. The only reliability comparisons discussed in the report were between TPH and total PAH, but these were only discussed in a general way. Individual PAHs and organic carbon normalization were apparently dismissed without any comparative analysis.

Data Interpretation and Use for Regulatory Decision-Making	
Do you agree with Ecology's conclusion that this approach is consistent with that of the WA SMS marine standards where the SQS and CSL were established as the lowest and 2nd lowest of the Apparent Effects Levels determined for a suite of bioassays?	I agree that it is consistent in spirit, but there is a much more limited suite of bioassays available for the freshwater database: essentially only 2 species and 3 endpoints-Hyalella survival, Chironomus survival and growth with sufficient data.
Given the types of historic sediment data available, is the TPH method the best available approach for assessing the overall effects of petroleum hydrocarbons? What is your answer based on (theory or empirical data)?	TPH method—is the analytical method sufficiently well-defined and consistent throughout the dataset? I am not aware of any data that suggests that same TPH (or total PAH) concentrations from different sources have the same toxicity. Unless a mixture approach is used that can assess the cumulative toxicity of the individual components (eg, PAH toxic units),treating individual analytes as separate estimators may be the best approach

Are the measures introduced in the model to assess covariance, coupled with other available statistical tests of covariance, sufficient to address the inevitable co-occurrence of chemicals in the field when developing chemical criteria?	The FPM, like all other empirical approaches, is assessing the toxicity of environmental chemical mixtures where covariance is assumed. The FPM helps to identify the best estimators of toxicity in the database, but it should not be assumed that the suite of values represents a unique solution.
Should any chemical classes be summed that were not summed in the model to reduce covariance (e.g. phthalates)?	Summing chemicals is appropriate when the composition of chemicals included in the sum is consistent, and, when applied, it is necessary that all chemicals included in the sum are measured.

Review Of Biological Criteria	
Is the proposed bioassay suite appropriately sensitive to protect the freshwater macro benthic community (i.e., typical taxonomic structure and functions such as a prey base to endangered species like salmon)?	Are one species of amphipod and one species of midge sensitive representatives of the freshwater macro benthic community (communities)? These bioassays represent endpoints that are currently well-standardized and have matching chemistry and toxicity data available, but no information was presented to assess whether the endpoints used are "appropriately sensitive" or representative of the freshwater macrobenthic community It would be preferable to have additional taxa and sublethal endpoints represented.
From your experience, are there other freshwater bioassays/species that provide consistent, reproducible and sensitive results that should be considered for developing biological criteria?	I am not aware of other species that have the same level of method standardization and testing , but USGS is developing freshwater mussel tests that should prove useful.
Are there issues you may be familiar with regarding running and interpreting these bioassays (e.g., problems associated with culturing animals, confounding variables that may warrant protocol modifications, choice of toxicant for positive controls, etc.)?	No information

What comments do you have on the appropriateness of the various growth endpoints under consideration nationally (e.g., dry weight, ash-free dry weight, length)?	My understanding is that ash-free dry weight is required for *Chironomus* (but not *Hyalella*) and either dry weight or length is an appropriate measure for *Hyalella*. Both growth endpoints should be converted to measures of total biomass, which combines mortality and growth into one endpoint.
Is there additional information on the minimum detectable difference for these tests that would assist in setting the SQS (SL1) interpretive endpoint?	Is there sufficient data in the freshwater database to evaluate MDDs?
What are the pros and cons of using two endpoints from the same bioassay (e.g., the Hyalella 28 day test has two endpoints commonly reported from the same test, mortality and growth). Are there ways to maximize the use of the combined results?	As above, I recommend using the biomass endpoint to combine mortality and growth, as these endpoints cannot be considered to be independent of each other. Alternatively, using a growth endpoint that represents either mortality or growth effects could be used as a single endpoint.

Review of WDoE Proposal for Freshwater Sediment Standards

by

G. Allen Burton

October 26, 2010

The Washington Department of Ecology (WDoE) is commended on another progressive step forward in developing tools to better manage aquatic ecosystem impairment. The current document and background materials are very useful and have a strong science base, as well as management base that justify the approaches being proposed. As with most efforts that are so complex, there are inherent issues that are raise some concerns from a scientific and transparency perspective. I do not address every question that was provided, but offer the following comments for WDoE's consideration. I have repeated the WDoE questions where appropriate and then given my responses in red.

- Do you agree with Ecology's conclusion that the use of sediment bioassays provides a credible basis for predicting adverse benthic effects that is consistent with current scientific information? *Yes. Well accepted, but in the same context, should only be done in a "weight-of-evidence" approach that utilizes multiple assessment approaches. Please acknowledge up front the limitations of SQGs as noted in Wenning et al 2005 (the SETAC Pellston conference book on SQGs). Among these limitations note the uncertainty associated with assigning causality between single chemical SQG exceedances when the empirical data is from sites with multiple chemical exposures.*
- Do you agree with Ecology's conclusion that multivariate statistical analysis provides a credible basis for characterizing the relationships between chemical concentrations and biological test results? *This statement is too extreme. They assist in the decision making process, but cannot be used alone- particularly given the inherent lack of causality relationships that are associated between the two.*

1. **Data Issues: The data used to develop the draft chemical criteria is summarized in Section 2.1 of the technical report.**

- Do you agree with Ecology's conclusion that the database provides sufficient data to support the development of statewide chemical criteria? That is, is the dataset sufficiently robust for the development of chemical values specific to:
    - Geographical coverage.
    - Coverage of different types of freshwater systems.
    - Numbers of paired chemistry and bioassay endpoints.
    - Number of bioassay species.
    - Number of acute and chronic tests (referring to test duration relative to life history).
    - Number of lethal and sublethal effects endpoints.
    - Percentage of hits and no-hits for each endpoint.

  The database is limited and needs to be augmented, from many perspectives including: geographically, chemically, concentration ranges, and biological effects (bioassay and indigenous). If one looks at the limited concentration ranges for some

chemicals it is obvious that the resulting frequency distributions are weak from a scientific and ecosystem relevance perspective.

- Do you agree with Ecology's conclusion that data was screened using criteria acceptable for the purposes of chemical criteria development? Specifically, refer to section 2.1.2 of the technical report which includes completeness of sediment chemical analysis, minimum number of detected analaytes, QA/QC of sediment chemistry and bioassay data, and elimination of chemicals that were not directly associated with toxicity.

  Some of the decisions seem arbitrary and contrary to typical, national, approaches. There needs to be increased transparency as to why these decisions were made. Why would you use 5 replicates vs. 4. Is WDoE's scientific vetting process superior's to ASTM or USEPA's? All of the screening criteria need a strong scientific basis – else you risk the perception that data were deleted to make it work for WDoE….

2. **Reliability Testing of the Chemical Criteria**: The state and federal agency representatives involved in this effort decided that it was preferred to use all available data to generate the criteria rather than to hold several datasets out of the process for validating the model. A validation study dependent upon new paired sediment chemistry and bioassay data would likely require years of data collection. Thus, reliability was evaluated with the dataset used to generate the values, looking at several reliability endpoints (refer to Section 2.3 Reliability Analysis in the technical report).

   - Do you believe that the approach used to evaluate the reliability of the criteria is consistent with current scientific methods and principles for validating criteria? *I am particularly concerned that sediment chemistry and bioassay sediment sampling were not co-located – given the massive spatial heterogeneity that exists in most sediment sites. Were samples split for chemistry and bioassays? If not – state it. Please do not ignore this huge uncertainty level…*
   - What comments do you have on the completeness and relative weight that should be given to the various reliability measures used to assess the results of the model and to compare it to other SQG sets? *Why are the criteria not compared to ERM/ERLs or PELs/TELs, etc.? These criteria have been documented to be useful and of similar accuracy to yours, so their comparison is absolutely essential. If you do not do it, someone else will and then report it in the peer-reviewed literature.*
   - Comparative reliability analysis was used to assess different ways to handle data with respect to number of issues. For example, the toxicity of petroleum compounds can be assessed using various approaches (TPH, individual PAHs, total PAHs, normalizations, etc.), and there were other similar decisions that needed to be made regarding the underlying data set (e.g., summing, normalization, inclusion of conventionals, splitting or lumping geographic areas). Were the reliability comparisons the most appropriate method to make these decisions or are there better methods that could be used? *I have a real problem with WDoE suggesting that their way is better than other national/international ways that have been better vetted through the scientific review process. TPHs have not been shown to be useful – as they are all over the place. WDoE is discounting AVS, TOC, grain size, Fe/Mn/Al as being useful. While none of these are*

useful at all sites, they CERTAINLY can be useful to help with the risk/hazard decision making process at many sites (see the peer reviewed literature). This is disconcerting and will do nothing but raise legal challenges and concerns.

3. Data Interpretation and Use for Regulatory Decision-Making:

- Do you agree with Ecology's conclusion that this approach is consistent with that of the WA SMS marine standards where the SQS and CSL were established as the lowest and 2nd lowest of the Apparent Effects Levels determined for a suite of bioassays? *Seems reasonable from a management perspective. The description of the Floating Percentile and balancing of Type 1 and 2 errors is excessive and in fact confusing. Its not that complicated. The approach is fine and the accuracy rates reported are similar to the other more commonly used SQGs.*
- Given the types of historic sediment data available, is the TPH method the best available approach for assessing the overall effects of petroleum hydrocarbons? What is your answer based on (theory or empirical data)? *No, see response above under number 2.*
- Are the measures introduced in the model to assess covariance, coupled with other available statistical tests of covariance, sufficient to address the inevitable co-occurrence of chemicals in the field when developing chemical criteria? *Well, yes and no. Each site is different and that should not be forgotten when making site-specific decisions.*
- Should any chemical classes be summed that were not summed in the model to reduce covariance (e.g. phthalates)? *Only if there is an adequate database (such as for the 16 PAHs) to justify.*

REVIEW OF BIOLOGICAL CRITERIA

- Is the proposed bioassay suite appropriately sensitive to protect the freshwater macro benthic community (i.e., typical taxonomic structure and functions such as a prey base to endangered species like salmon)? *It's the best you can do at present. Consider adding snails and mussels in the future as they are important and sensitive. Hopefully these document and related policy will be reviewed every couple of years as the science advances.*
- From your experience, are there other freshwater bioassays/species that provide consistent, reproducible and sensitive results that should be considered for developing biological criteria? *See previous response.*
- Are there issues you may be familiar with regarding running and interpreting these bioassays (e.g., problems associated with culturing animals, confounding variables that may warrant protocol modifications, choice of toxicant for positive controls, etc.)? *Two primary concerns mentioned under 1 and 2 above (5 vs 4 reps, sediment chem./bioassay sample splits)*
- What comments do you have on the appropriateness of the various growth endpoints under consideration nationally (e.g., dry weight, ash-free dry weight, length)? *Always do the least variable growth measure to better detect site differences (higher discriminatory power).*

- Is there additional information on the minimum detectable difference for these tests that would assist in setting the SQS (SL1) interpretive endpoint? *None that I know of.*
- What are the pros and cons of using two endpoints from the same bioassay (e.g., the Hyalella 28 day test has two endpoints commonly reported from the same test, mortality and growth). Are there ways to maximize the use of the combined results? *Both fine, as one is acute and one is chronic. Hopefully your chronic database will grow.*