July 3, 2014

Comments on UTDWQ 2014 Integrated Report

David C. Richards, Ph.D.
OreoHelix Consulting
P.O. Box 996
Moab, UT 84532

General Comments

Many of the criteria/analyses used in the 2014 Narrative were developed specifically as first round 'screening' tools, particularly O/E bioassessment. These assessments should not be used for anything other than their intended use; initial screening. They were never intended for monitoring or any scientific analysis. Assessments by their very nature are a critical link between science and managers. Indeed managers often depend on scientists to develop easily understandable measures of ecosystem health and scientists concerned with the ecosystems that they study are civilly obliged to produce the most meaningful, state-of-the-science assessments to managers. However, assessments are not science and should not be used as a substitute, although assessment results can sometimes be used to develop scientific hypotheses for further studies. In addition, and contrary to current management agencies' agendas, assessments are not valid monitoring tools, primarily because of their poor discriminatory power and lack of ability to measure anything less than very large changes in ecosystem health and they should not be used as such.

<u>Example of Bioassessment tool that is too simplified for assessment, monitoring or scientific investigation</u>
For example, the O/E biocriteria method produces a single value that was designed for a single group of organisms (macroinvertebrates) and their assumed response to a combination of many environmental stressors. O/E models first determine which group a 'test' stream is to be included in prior to O/E calculation (AU grouping methods are also subject to scrutiny and will be discussed further). Once the grouping of the stream is assigned, then O/E is calculated. Again, macroinvertebrate taxa in O/E are responding to generalized, accumulative types of stressors. However, it is well known that individual macroinvertebrate taxa respond differently to a wide array of environmental stressors. That is one reason they have evolved as separate species. Macroinvertebrate taxa and certainly entire aquatic ecosystems are not simple, single- dose response interactions. Aquatic ecosystems are extremely complex with interactions occurring at multiple levels, starting from individual species up to interaction with the entire ecosystem. Each of the hundreds of species in an aquatic ecosystem has a niche that can be defined as a

Hutchinsonian *N*-dimensional hyper- volume, not just one, two or three dimensions, but many. Multiply these multidimensional niches of each species by interactions with all the other species (competition, predation, parasitism, mutualism, facilitation, etc.), trophic level interactions, functional feeding groups, interactions between aquatic and riparian ecosystems, stream connectivity (which include metapopulation dynamics and genetic interactions and which were part of Karr's definition of 'biological integrity' but are now completely ignored by water quality management agencies), natural spatial and temporal variability, anthropomorphic impacts, etc. etc. and any stream ecologist would shake their head from side- to- side in complete disbelief that the health of a stream could be summed up to one subjectively determined number (score), even if it means making things simple for managers to easily comprehend. One single O/E score cannot possibly hope to capture the health of an entire aquatic ecosystem. It is just not possible. Thus the reliance of O/E to assess water quality condition and whether a water body supports or doesn't support its designated use often results in an injustice to water users and regulators, particularly those who are responsible and dedicated to maintaining water quality. It also does a disservice to the very waters of UT, themselves and the biota that reside in them. Utah's rivers and streams are much too valuable and a treasure to all its residents to use flawed, highly simplified, or limited number of metrics. The reliance of O/E by UTDWQ also can reflect poorly on the very agency that citizens of UT have entrusted to protect their waters.

Type I and II Error and Sample Size

Thoughout the Narrative, DWQ uses Type I error rates to support its decisions concerning water quality conditions, i.e. "Is there a problem?" and "How extensive is the problem". Type I, a classical frequentist's statistic, tests the null hypothesis of no effect vs. the alternative hypothesis of an effect, in this case; no impairment vs. impairment. Type I error, measured as alpha or 'p' (1-alpha), occurs when the null hypothesis is actually true but was rejected as false. That is, there truly was no impairment but the conclusion was that there was impairment (also known as a false positive). Type I error is often illustrated by the story of 'crying wolf' when there actually wasn't a wolf. If one cries wolf too many times, no one would believe them when the wolf actually showed up. By focusing on Type I error, particularly with small sample sizes, too often UTDWQ may have ended up reporting impairment when there really was no impairment This could add a considerable workload and expense down the road for UTDWQ or other managers trying to determine if indeed there was impairment which then must commit to an expensive TMDL or other restorative efforts or commit additional resources necessary for delisting a water body. It can also result in doubt of DWQ's ability to truly detect impairment.

It has been suggested that this approach may be "proactive, precautionary, or erring on protection, etc." This may be true, particularly given the small sample sizes used in the assessments. The number of samples used in an assessment determines the

level of ability to detect changes. Small sample sizes typically only allow for detection of large differences, whereas large sample sizes can detect small differences in assessment tests, although in some cases small differences may be ecologically irrelevant. *However, by limiting sample size to very small levels, individual data points will have undo influence, particularly outliers. These few data points could prompt UTDWQ to conclude impairment when in fact there was no impairment. If more data were used, a single data point would have less of an effect on the conclusion.* An example would be if three samples were collected in the same location and a metal such as aluminum was found to be in exceedance of standards in one of those three samples (33% of the samples) vs. the same sample was found to be in exceedance but one thousand samples were collected (0.01% of the samples).

Type II error, on the other hand, is a more precautionary type of error that may have been of better use by UTDWQ for determinations. Type II error (B) is the probability of concluding there wasn't impairment when in fact there was (not crying wolf, when it actually it was there). This is of major concern because DWQ may be evaluating UT waters incorrectly by not detecting true impairment and possibly allowing impairment to continue. Significantly and ecologically meaningful Type II error levels often require substantially more data points. However, by collecting additional data and using existing data from all available sources, assessments that incorporate Type II error could save time and money in the long run.

The reliance on Type I error evaluation allows DWQ to say, "oops we thought there was an impact but on closer evaluation it looks like there wasn't". "Guess all that money spent to fix things wasn't necessary after all". In addition by using only a few samples, Type II error would likely result in the inability of the assessment to detect a true impact, which could be more harmful than Type I error. Again, more samples should be included in an assessment.

The term "Biological Composition" used throughout the Narrative

Not sure what 'biological composition' means, but how ever UTDWQ defines 'biological composition'; O/E does not measure it. O/E purportedly only measures taxa richness, not composition. Community or assemblage composition of course is not the same as taxa richness.

Comment X. Introduction Page 6: First paragraph last sentence: "Even completely subjective….(fish kills..)

Not sure how a fish kill can be considered less subjective than the other more preferred methods of evaluation used in the narrative. I guess if one was to say, "all those fishies look dead to me" (i.e. subjective) as opposed to going over and kicking a few and smelling them (objective) and concluding, "yep, they are dead" would count. In reality all of the methods used and decisions made in this narrative are subjective, including decisions based on statistical tests. This is the inherent nature

of every conclusion based on statistical inference. A decision to use a Type I error alpha level of 0.05, 0.10, or any value is subjective. Why not alpha = 0.06? or 0.04? The decision to be consistent with the choice of an alpha level for every test (e.g. 0.05) and for every circumstance is also subjective. An alternative would be to examine each test (criterion) and its resultant alpha level and then making a conclusion based on their own merits or importance. This is particularly critical due the economic and ecological importance of a decision to list an AU as impaired or not. Fish kill observations are perhaps the most objective of all the decisions used in this narrative. If a fish kill is observed, then the stream is impaired. However, determining the cause of the kill would require further investigation.

Category 4C page 8, first sentence

"Category 4C: The impairment is not caused by a pollutant: Assessment units are listed in this subcategory if the impairment is not caused by a pollutant (e.g., habitat alteration, hydromodification)."

Comment: Many of the streams listed as "not-supporting" in the Narrative should have been in this category and not in the not-supporting category.

"·Quantitative biological assessment results for streams and rivers are statistically different than the reference site conditions."
Comment: Assessments could be statistically different than reference conditions because of many factors other than impairment (see example below).

Page 34, First Sentence: "E is then calculated as the sum of all taxa Pc s that had a greater than 50% chance of occurring at a site given the site's specific environmental characteristics."

It appears that O/E development requires that taxa have a probability of occurrence of > 0.50 in (reference) streams to be part of the model. If this interpretation is correct, then an unknown number of taxa are automatically removed from consideration in O/E. This unknown number of omitted taxa could be very large depending on which group the reference streams are placed under. It appears that only the ubiquitous, common, cosmopolitan, tramp species are used as 'observed' taxa; taxa that are likely to be tolerant of a wide range of environmental variables and not likely responsive to stressors. For example *Baetis* sp. (mayflies) have a probability of occurrence for a reference group of 0.95 (expected to occur). Obviously, certain *Baetis* sp. are cosmopolitan (e.g. *Baetis bicaudatus*) and which have a wide range of environmental tolerances. *B. bicaudatus* also occur in streams in less than reference condition (e.g. Mill Creek, SLC at confluence with Jordan River). Thus baetid mayflies as a group and particularly *B. bicaudatus* are very poor indicators of water quality. As far as a test sample observed inclusion or exclusion of *B. bicaudatus*, or any other taxon, it would be entirely dependent on whether a complete census of the entire reach of stream under consideration was conducted

or not. If a census of the entire reach of concern was not conducted and only composite sampling and then laboratory subsampling was conducted, it would be impossible to know if that taxon was truly present or absent. Low abundance of a taxon and hence its omission from observed status does not mean extinction of that taxon from a water body. It simply means it was not observed. In addition, a taxon could occur at 90% of the total assemblage abundance in many streams (i.e. be locally abundant within some streams), but have a probability of occurrence of < 0.50 in all reference streams and therefore, excluded from the O/E model. The O/E model assumes that macroinvertebrate assemblages in all streams in a group are similar and individual taxa occur at equal abundances and none are unique or whose populations are not dynamic. However, the IR mentions that samples should be conducted or evalutated every three years or so to account for natural variability in taxa abundances. Again, all of these assumptions are not likely. 1) A test stream may fit poorly into its designated reference group and its macroinvertebrate assemblage may be quite a bit different than reference, 2) taxa don't occur at equal abundances within any stream and therefore have unequal probabilities of detection using DWQ methods, and 3) a major factor in the population dynamics of a taxon is its generation time. Taxa with short generation times (e.g. midges; several weeks to a few months) have greater variability in abundances due to environmental conditions than do taxa with longer generation times (e.g. large stoneflies, 1 – 3 years). If test samples aren't conducted during a similar point in a taxon's natural population abundance cycle, erroneous conclusions about 'observed' status will be made.

As stated in the IR, O/E ignores many taxa; many of which may be rare, uncommon, cryptic, or even may be very common in many streams. This could include federally listed threatened and endangered species which most U.S. citizens support the protection of. UTDWQ appears to be ignoring an unknown number of taxa, not assessing their status (i.e. biological integrity), and differing to USFWS to deal with them under the Endangered Species Act (if they ever make it to that list). Also, rare and uncommon taxa are much more likely to be indicators of water quality and should be the focus of water quality assessments rather than cosmopolitan taxa. In addition, ecologists are now well aware that rare and uncommon taxa often have disproportionally greater influence on ecosystem function than common taxa. For example, the salmonfly, *Pteronarcys californica* may have a probability of occurrence <0.50 in the stream group but when it occurs in large abundances in some streams it has a disproportionately large influence on the entire ecosystem, from reducing CPOM to FPOM all the way to a being a critical food item for breeding birds when it occurs as an aerial adult. Other less common taxa such as the stonefly *Yoroperlla* sp. or caddisfly *Helicopsyche borealis,* etc. are likely excluded from the models. These examples and many other 'uncommon' taxa are uncommon for several reasons: because they have limited geographic distributions or distributions within a region or streams, limited range of environmental tolerances and conditions, or are less tolerant to human disturbance than the cosmopolitan taxa used in the model. The IR acknowledges the problems with 'rare' and 'uncommon' taxa but it is likely that far

too many taxa are considered 'rare' or 'uncommon' in the O/E models and are done so if favor of making the O/E models function.

How many or what proportion of taxa can occur in < 50% of the streams and were excluded from the O/E model? If O/E represents local extinctions then the status of these taxa were not included in the estimate and O/E could be grossly underestimating local taxa extinction. Again, O/E would not be quantifying loss of biodiversity except in the crudest sense and if it does it will be only for ubiquitous taxa.

Section: RIVER INVERTEBRATE PREDICTION AND CLASSIFICATION SYSTEM (RIVPACS) MODELS\Page 32
3rd paragraph second sentence: "In essence, O/E quantifies loss of biodiversity."

Comment: **No it does not.** O/E does not quantify loss of biodiversity. It may on occasion, but it is unknown from the assessment if loss of biodiversity (taxa richness) actually caused the change in O/E values. Particularly when fixed count subsampling methods are used in the taxonomy lab. It could very well be that biodiversity hasn't changed but that one or more taxa may have happened to become more or less abundant (i.e. change in evenness). If used at all, O/E should more appropriately be used as a rudimentary measure of 'evenness'. Please see hypothetical example below.

"Despite the mathematical complexities of model development, O/E is easily interpreted as it simply represents the extent to which taxa have become locally extinct as a result of human activities. For example, an O/E ratio of 0.40 implies that, on average, 60% of the taxa have become locally extinct as a result of human-caused alterations to the stream"

Comment: Again, this is likely not true (see example below).

In addition, the statistical methods (models) that went into the development of RIVPACS O/E model have associated error or variability. For example, cluster analyses that were used to develop reference groups have associated error rates and there are many cluster analysis methods available, each potentially resulting in a different set of reference groups. DWQ likely used the most appropriate cluster method based on either cluster model comparisons or best professional judgment or both. However, there still are error rates associated with the best method used. The probability of occurrence of a taxon in a reference group also has inherent uncertainty or error. RIVPACS O/E models are 'models within models' each of which contributes uncertainty either additively or multiplicatively. These error rates need to be taken into account and reported in the IR.

Major problems with O/E

Natural Variability and Sampling Error

1) Natural variability (e.g. annual, seasonal, and year- to- year variability in physical conditions and macroinvertebrate abundances), 2) within stream variability (riffle to riffle or riffle to other type of habitat e.g. riffles tend to have more taxa than pools), 3) field sampling error (e.g. estimating the 1 sq. ft. area needed to be sampled), 4) sample processing error (e.g. proper preservation, handling, and storage), and 5) laboratory error (rolling up of taxa, different levels of taxonomic QA/QC, etc.); while hopefully kept to a minimum can add up to the likelihood of erroneous O/E scores far greater than the 0.01 level used by UTDWQ to conclude fully supporting vs. non- supporting. For example, the UTDWQ O/E scores of 0.83 or 0.78. It appears that in an O/E model with for example, 100 taxa, the assumed loss of only one taxon could result in a change in use support status even though it could have been due to natural variability or sampling or modeling error.

Fixed Count Subsampling Error

Composite samples of eight, 1 –sq. ft. kick samples recommended/endorsed by UTDWQ can often have large number of individual organisms, sometimes > 10,000 individuals. To reduce the amount of cost and effort in processing this large number of organisms and to standardize samples across regions, UTDWQ and O/E models typically use laboratory produced 500 organism subsamples. Data from these processed samples are then entered into O/E resulting in scores that are assumed to represent taxa richness and/or "the percentage of taxa that have become locally extinct as a result of human- caused alterations to the stream". Again, this is most likely an incorrect conclusion as illustrated by an example below.

Hypothetical example of UT DWQ O/E miscalculation and false conclusion

The following is a hypothetical example of the O/E fixed count subsampling problem.

Methods: Two composited samples from eight, 1-sq ft. kick samples were collected at the same location but in different years. Sample 1 was collected the year prior to Sample 2. Both samples were collected in riffles and 'other' acceptable habitats (e.g. runs, banks, etc.). There were ten taxa collected in both samples; two mayfly (Ephemeroptera) taxa, two stonefly (Plecoptera) taxa, one caddisfly (Trichoptera) taxon, two midge (Chironomidae) taxa, and one taxon from the following groups, snails (Gastropoda), scuds (Crustacea), and segmented worms (Oligochaeta). In the first sample, the number of individuals for each of the ten taxa was 1000. In the second sample there were substantially more mayflies and stoneflies than the first sample and less individuals of the other six taxa (Table 1). In both samples the total number of individuals was equal; 10,000. This number of individuals is not unusual for composited samples from stream systems in UT. Results of the mean numbers of individuals of each taxon for each of the two samples using a 500-organism subsample method are in Table 1.

Table 1.

			Composite Sample Number of Individuals		500 count subsample Mean Number of Individuals	
			Sample 1	Sample 2	Sample 1	Sample 2
Ephemeroptera	Mayfly	*Serratella* sp.	1000	2525	50	126.25
		Drunella sp.	1000	3000	50	150
Plecoptera	Stonefly	*Zapada sp.*	1000	2000	50	100
		Capnia sp.	1000	2000	50	100
Trichoptera	Caddisfly	*Rhyacophila sp.*	1000	400	50	20
Gastropoda	Snail	*Physa sp.*	1000	15	50	0.75
Chironomidae	Midge	*Chironomus sp.*	1000	15	50	0.75
		Tanypus sp.	1000	15	50	0.75
Crustacea	Scud	*Hyalella sp.*	1000	15	50	0.75
Oligochaeta	Worm	*Tubifex sp.*	1000	15	50	0.75
Total # of individuals (organisms)			10,000	10,000	500	500
Total Taxa			10	10	10	5

Any mean values < 1.0 in the 500 count subsample (Table 1) indicates that on average the taxon occurred less than once in the 500 count subsample and was therefore, never observed or counted. Any mean values > 1.0 indicates the taxon occurred in the 500-count subsample and was observed and counted.

Results: The mean total number of taxa counted and reported from the 500-count subsample in Sample 1 was **ten** and the mean total number of taxa in sample 2 was **five**. This represents a 50% difference in total taxa reported, even though there was actually the same number (10) of taxa collected in the original samples.

Conclusions:

The conclusion using UTDWQ O/E criteria would be that:

50% of taxa became extinct from when Sample 1 was taken to when Sample 2 was taken due to human impact. Therefore, the stream is not supporting its designated use.

The conclusion of a stream ecologist would be that:

Biodiversity may not have changed from year- to -year and production (total number of individuals) may not have changed, as well, but cannot tell using results from a fixed subsample method. If the entire samples were analyzed, then biodiversity and production did not change. Indicator taxa that often represent good water quality (mayflies and stoneflies) increased by 2 to 3 times, and those typically considered poor water quality indicators decreased by almost 67 times. However, most stream ecologists disagree with bioassessment programs that suggest that all midges, snails, crustacean, and worms should be classified as poor water quality indicators and caution should be applied to this statement.

The conclusion of the stream ecologist would also be that it appears that water quality improved and could have been due to natural variability or given additional data, likely improved because of decreased water temperature or conditions that favored mayflies and stoneflies, particularly stoneflies in the functional feeding group, shredders. Increased shredder abundance was likely due to increased riparian cover. The scraper snail taxon *Physa* sp. may have decreased in abundance due to less light from increased riparian cover. Whatever those humans did (e.g. increased riparian cover which may have decreased temperature, increased allochthonous production, and decreased autochthonous production); Keep up the great work!

Using fixed count subsample method could have resulted in just one single taxon not being counted if it occurred at low abundances in the stream and thus lowering the O/E score from supporting to not supporting even if it was present in the stream.

Jordan River Question
Was O/E conducted for all sites on Jordan River or just those that resulted in not-supporting? If O/E wasn't conducted in these AUs then decisions were based on only one type of measure (line of evidence)(e.g. chemical, etc.). Also, was there a reference stream to compare the Jordan River to and what was it?

Conclusion

1. Assessments are a simplified tool to aid managers in their decisions, nothing more. Assessment methods should not be used to monitor water quality and their results should not be interpreted as scientific evidence. Assessments should also not be considered as a substitute for good science.

2. Results of assessments that rely on very small sample sizes should be interpreted with extreme caution. Many streams listed as 'not- supporting' should likely be relisted as Category 3, insufficient data and others should be listed as 4C.

3. Biological integrity and ecosystem function, as defined in the Clean Water Act, can not justifiably be summarized into one score because these concepts are extremely complex, there is prolific natural variability, and error rates associated with sample collection, sample processing, and final score calculation may not simply be additive but are likely compounded with every step in the assessment procedure. UTDWQ should not claim that O/E is a measure of change in biodiversity (taxa richness) because other factors including those discussed in these comments, likely effect scores. O/E scores should be interpreted with extreme caution and not used as a primary tool in assessments.