

APPENDIX E

Threshold Development – Technical Underpinnings

Water Quality Control Division
August 2010

The Multi-metric Index (MMI) assesses biological condition on a scale of 0 to 100. The numeric characterization is analogous to measuring the concentration of a pollutant, but the number alone does not define what constitutes attainment (or impairment) of the use. Establishing biological thresholds is largely a statistical endeavor, but it must be preceded by a policy decision that defines use support on the basis of site characteristics. Once the basis for use support is established, appropriate statistical methods can be applied to determine the thresholds defining biological condition consistent with use support. Finally, there are implementation issues that the Division anticipates regarding differences between Class 1 and Class 2 waters, as well as a concern about protection of high-scoring waters.

Defining Use Support

The definition of use support is central to the derivation of biological thresholds because it controls membership in the group of sites that are in attainment. Once group membership is established, the statistical properties of biological condition (MMI) can be determined for sites in the group. Sample size, central tendency, and variance will shape the development of thresholds.

It is common practice to tie the definition of use support to the concept of reference conditions. In EPA guidance documents, for example, “use support” is equated solely with the technical definition of “reference conditions” (i.e., the basis for tool development). Moreover, some percentage of low-scoring reference sites is excluded. However, when use support is equated with reference conditions, the range of policy options can be adjusted only by the relatively trivial method of varying the percentage of reference sites excluded.

Defining reference conditions was an integral part of bioassessment tool development because it established one end of the spectrum for biological condition. Because selections are made before the tool is developed, there is a presumption that selection criteria will yield sites with good biological condition. Generally, this means selecting sites where anthropogenic stressors are minimal or absent. There is little doubt that minimally-disturbed¹ reference sites can serve the *technical* purpose of bioassessment tool development because they clearly represent one end of the spectrum for biological condition.

It can also be argued that stringent selection criteria, which minimize potential stressors, yield only those sites most easily associated with use support. Higher stressor levels may also be consistent with use support, as is clearly acknowledged in the TALU condition gradient, but it

¹ Minimally-disturbed in Colorado is likely to include atmospheric deposition of some pollutants (e.g., nitrate and mercury of anthropogenic origin).

would be hard to know this without having first developed the tool. Once the tool is available, it becomes possible to assess biological condition independent of labels.

The point to be emphasized is that, at least for Colorado, the Division believes reference sites represent a subset of all sites at which the aquatic life use is supported. Whether that subset constitutes 50% or 80% or 95% of sites supporting the use is guided to a large extent by policy that the Commission will set. The concept of reference condition, which was essential for bioassessment tool development, is not indispensable for reaching conclusions about use support.

Biological Thresholds Derivation

Development of biological thresholds falls largely in the policy realm, albeit with a strong reliance on statistical tools. The goal is a characterization of biological conditions demonstrating support of the aquatic life use. Because policy options are both enabled and constrained by this characterization, the definition of use support represents a critical exercise of policy prerogative.

The Division believes the Commission should be able to consider a range of policy options in setting biological thresholds. These options, which are described below, may include biological conditions that are more expansive (or more restrictive) than those circumscribed by the technical definition of reference conditions. The key elements for option development include the basis for membership in the group representing use support and the characteristics of the MMI distribution for that group.

The statistical approach recommended by EPA involves the application of interval and equivalence tests for the purpose of locating attainment and impairment thresholds. The method is described in Kilgour et al.² It begins with a definition of the normal range, which establishes the acceptable range of biological condition within the group of sites representing use support.

Normal Range

According to Kilgour et al., the normal operating range is “typically defined as the range of values enclosing 95% of the population..., regardless of the discipline.” Previous EPA guidance bears superficial resemblance to the normal range concept in that a low percentile of the reference set is often used to set the impairment threshold. Beginning with Ohio’s work on biological thresholds, the 25th percentile of reference was set as the threshold for impairment. However, EPA guidance frames the choice somewhat differently by stating that the “actual percentile chosen ... is arbitrary and represents the amount of uncertainty that a monitoring program can tolerate.”³ EPA’s arbitrariness in selecting a percentile is evident in the range of values that has been applied in other states: 2.5% to 25%. No rationale is available to explain the preference for one percentile over another. Moreover, the absence of a rationale is puzzling in

² Kilgour, BW, KM Somers, and DE Matthews. 1998. Using the normal range as a criterion for ecological significance in environmental monitoring and assessment. *Ecoscience* 5(4): 542-550.

³ Barbour, MT et al. 1996. *Biological Criteria – Technical Guidance for Streams and Small Rivers*, Revised Edition. See <http://www.epa.gov/bioindicators/pdf/EPA-822-B-96-001BiologicalCriteria-TechnicalGuidanceforStreamsandSmallRivers-revisededition1996.pdf>

view of the primary goal, which is to determine “whether an individual site is a member of the least-impaired reference population.”⁴

Given the level of effort that Colorado has invested in selecting reference sites, application of the normal operating range (central 95% of the distribution of MMI scores; Figure 1) seems appropriate. Nevertheless, in deference to concerns raised by EPA Region 8, an option also is presented with a narrower operating range (central 90%).

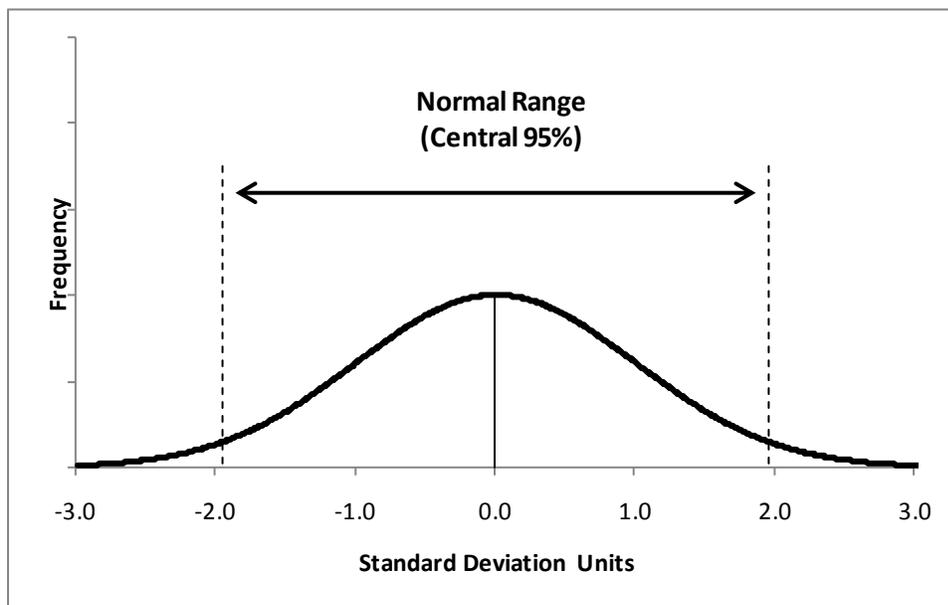


Figure 1. Graphical representation of the “normal range” based on the central 95% of the distribution (bold line represents the normal frequency distribution). The dashed vertical lines signify the boundaries of the central 95% at $\pm 1.96\sigma$.

Statistical Basis for Thresholds – Interval and Equivalence Tests

The Division is proposing to use MMI scores as the numeric biological thresholds that determine whether biological condition at a particular site is impaired or in attainment of the use. Establishing those criteria is a statistical exercise that incorporates policy considerations related to the definition of use support, as well as the risk of being wrong about concluding that a site is impaired or in attainment. The proper framing of the statistical question is, of course, central to the conclusions that can be drawn.

The traditional statistical approach tests the hypothesis that the site in question is a member of the group of sites with MMI scores representing use support (e.g., reference conditions). A test of this nature is straightforward and provides a clear-cut conclusion regarding the hypothesis (i.e., reject or not). The alpha value for the test defines the risk of mistakenly concluding that a reference-quality site is impaired (usually, $\alpha=0.05$, which means 1 chance in 20 of false rejection). However, the results would not reveal whether an ecologically-important difference is present, or even if one is detectable. Moreover, the power of the test may be low, meaning

⁴ Kilgour et al. 1998

that there would be a relatively high risk of mistakenly concluding that an impaired site was in attainment.

An alternative statistical approach is possible when the ecologically-important difference (also the critical effect size) is defined in advance. Although it might be difficult to decide in advance that a difference of 10 MMI would be considered ecologically-important, it would be both clear and simple to define the difference in terms of normal range (see above). When the normal range is defined in terms of standard deviation units, the concept is readily transferable among the biotypes and can be adjusted easily if group membership is augmented with additional sites (see below).

The alternative statistical approach, which is recommended by EPA, sets thresholds based on interval and equivalence tests. For these test, the observed difference between the trial MMI and the mean of the use support group is compared to the “ecologically important effect size”⁵. The interval test evaluates the hypothesis that the difference between the trial MMI and the group mean is smaller than the effect size; if it is rejected (i.e., the difference is significantly larger than the effect size), the trial site would be considered impaired. The equivalence test evaluates the hypothesis that the difference between the trial MMI and the mean is larger than the effect size; if it is rejected (i.e., the difference is significantly smaller than the effect size), the trial site would be considered in attainment.

Advantages of this statistical approach can be better appreciated by stepping through possible outcomes. For example, if the interval test is rejected, the outcome is simple: the site is impaired. Failure to reject the interval test, on the other hand, indicates no significant impairment, but that outcome does not necessarily mean that the site is in reference condition. The status of the site can be clarified by the equivalence test, rejection of which indicates that the site is probably in reference condition. If neither hypothesis is rejected, the trial value falls in a gray zone that could be interpreted as *possibly impaired*.⁶

⁵ Kilgour, BW, KM Somers, and DE Matthews. 1998. Using the normal range as a criterion for ecological significance in environmental monitoring and assessment. *Ecoscience* 5(4): 542-550. The Division has followed this paper in setting the effect size to correspond to the normal operating range (1.96 standard deviation units on either side of the mean encompasses the central 95% of the distribution) and in adopting a 5% level of significance for hypothesis testing.

⁶ Bowman, MF and KM Somers. 2006. Evaluating a novel Test Site Analysis (TSA) bioassessment approach. *Journal of the North American Benthological Society* 25(3): 712-727. A useful graphical representation of the hypotheses appears in this paper, which also outlines implementation of the computational procedures for Excel spreadsheets

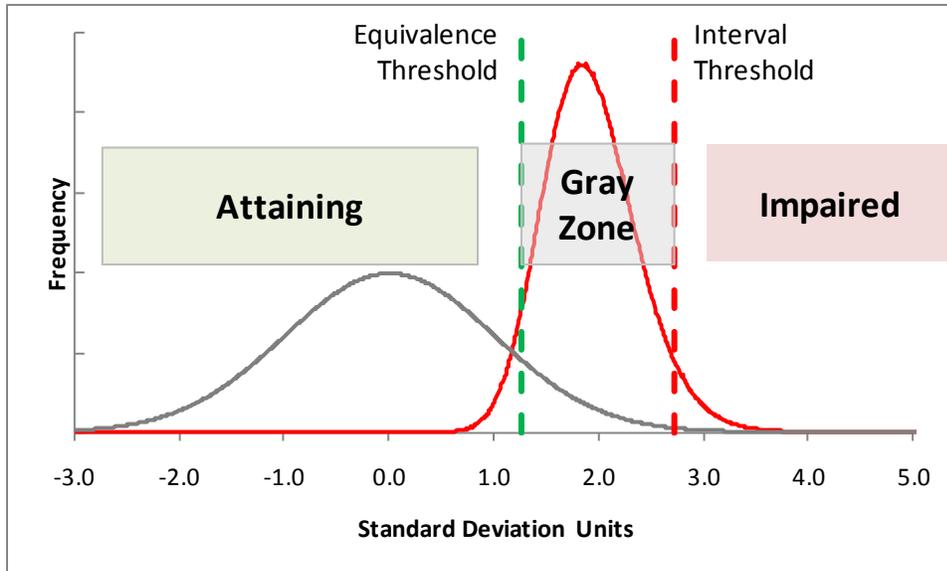


Figure 2. Graphical representation of thresholds derived from interval and equivalence tests. The distribution from which the normal range is derived is bell-shaped curve centered on 0.0 std deviation units. The non-central F distribution is shown as the taller curve centered on the Gray Zone. Equivalence and interval thresholds, which are shown as vertical dashed lines, divide assessments into three categories – attaining, gray zone, and impaired – as explained in the text.

Group Membership

Within each of the biotypes, some sites have been placed in the reference category, and some have been placed in a stressed category. The reference and stressed categories were important for tool development, as mentioned previously. However, a somewhat different perspective is appropriate for biological thresholds development. Biological thresholds focus on use support, which encompasses, but may not be restricted to, reference conditions.

Consistent with EPA guidance, reference conditions are used to define use support for one set of options developed by the Division. A total of 133 reference sites were identified through a rigorous screening process that included evaluation of human disturbance using GIS coverages, to produce the candidate reference list, and evaluation of aerial photos, to produce the final list. Reference sites were assigned to biotypes based on community composition (using cluster analysis as described in MMI development). For sites other than reference, biotype membership was based on predictive relationships. For each of the three biotypes, biological condition scores (MMI) were tested for normality and screened for outliers. All were normally distributed, and a single outlier⁷ was excluded from biotype 1 (Figure 3).

⁷ The Division has applied Rosner's sequential procedure to test for outliers because it handles the problem of masking when outliers may be close together. See: Gilbert, RO. 1987. Statistical Methods for Environmental Pollution Monitoring. Wiley, New York. 320p.

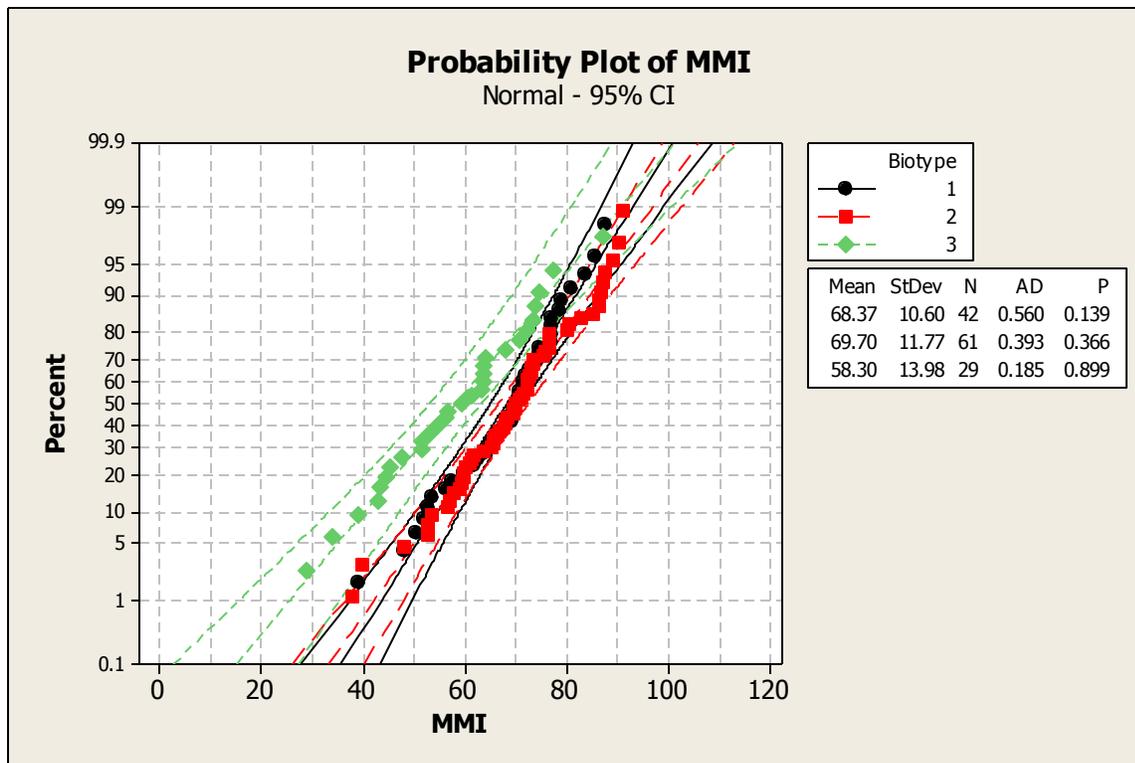


Figure 3. Probability plots of MMI scores at reference sites in each of the three biotypes. Biotype membership was determined by cluster analysis. One outlier was removed from biotype 1, as explained in the text.

The Division also has defined a group that expands membership beyond the reference sites. The added sites have human disturbance scores comparable to those measured at reference sites. Scores for human disturbance factors for all sites were derived from GIS coverages. Within each biotype, the range of scores recorded for reference sites was used to screen all non-reference sites. Any non-reference site for which scores for all disturbance factors were within the ranges observed for reference sites was added to the expanded set. The selection criteria are shown in Table 1.

Disturbance Factor	Units	Biotype 1	Biotype 2	Biotype 3
Irrigated Agriculture	% of WSA	<15.2	0	<36
Dryland Agriculture	% of WSA	<6.3	0	<49
Urban	% of WSA	<0.44	0	<0.31
Permitted Point Sources	#/km ²	0-1	0	0
Diversions	#/km ²	0-10	0-1	0-5
Road Density	Mi/mi ²	<3.6	<1.2	<2.8
Abandoned Mines	#/km ²	<0.6	0	<0.014
Oil & Gas	#/km ²	0-4; NA	--	0-1; NA
CAFO	#/km ²	0; NA	--	--
Sites added to group	#			

Table 1. Threshold values of human disturbance factors used to define membership in the Expanded set. WSA= watershed area.

The premise for the expanded set is that it defines a group with relatively low human disturbance scores, all of which are comparable to scores at reference sites. Consequently, sites in the expanded group are just as likely to support the use as the reference sites. Distributions of MMI scores for reference, expanded, and all sites are shown for each biotype (Figure 4-Figure 6).

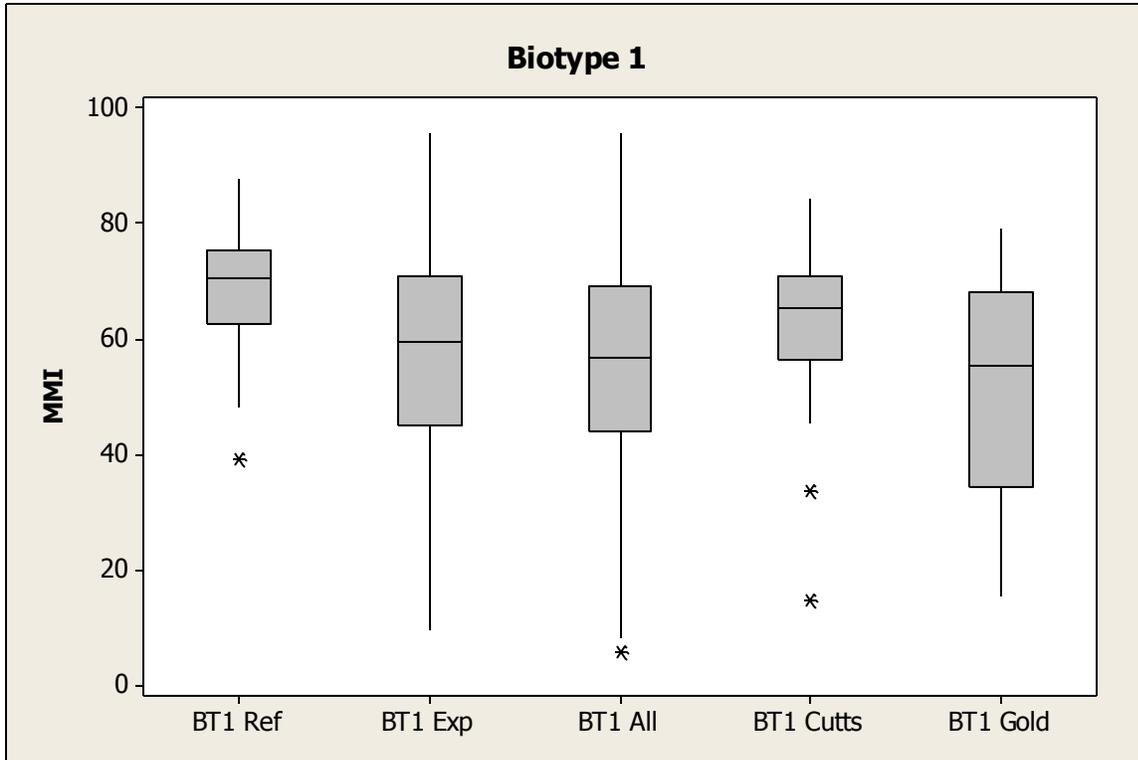


Figure 4. Boxplots of MMI scores for sites in biotype 1 aggregated by group membership (Reference, Expanded, All) or associated with Designated Cutthroat Trout Habitat or Gold Medal fisheries.

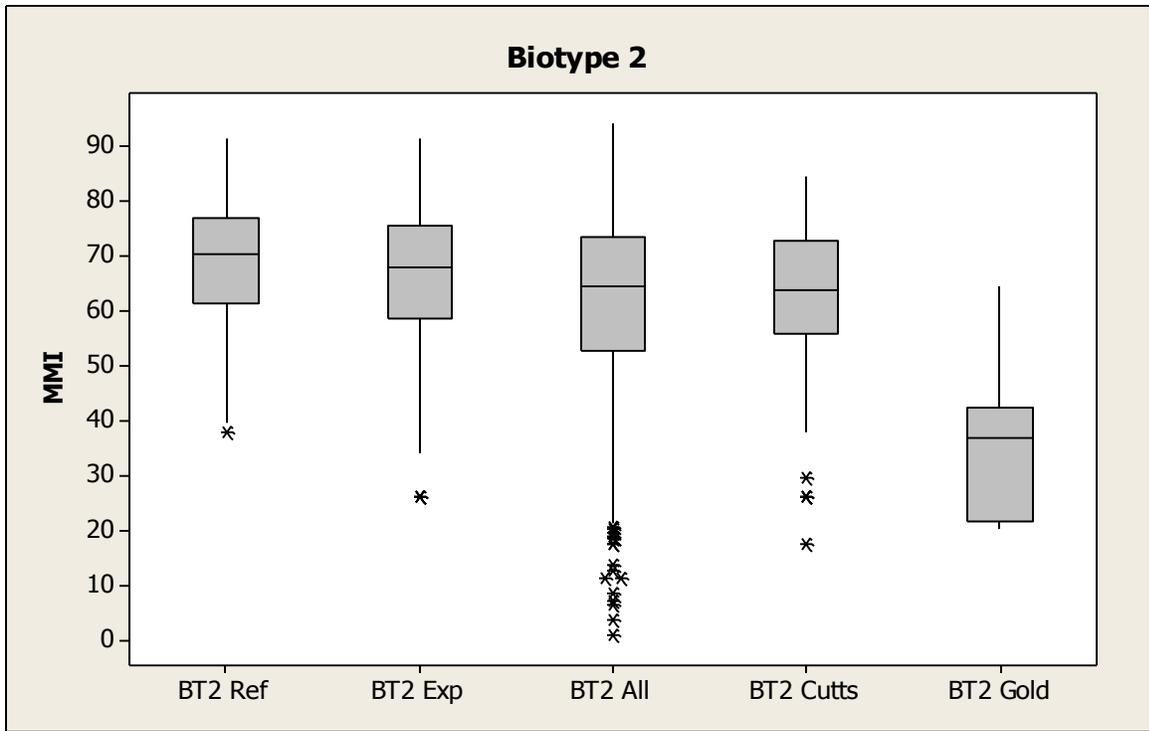


Figure 5. Boxplots of MMI scores for sites in biotype 2 aggregated by group membership (Reference, Expanded, All) or associated with Designated Cutthroat Trout Habitat or Gold Medal fisheries.

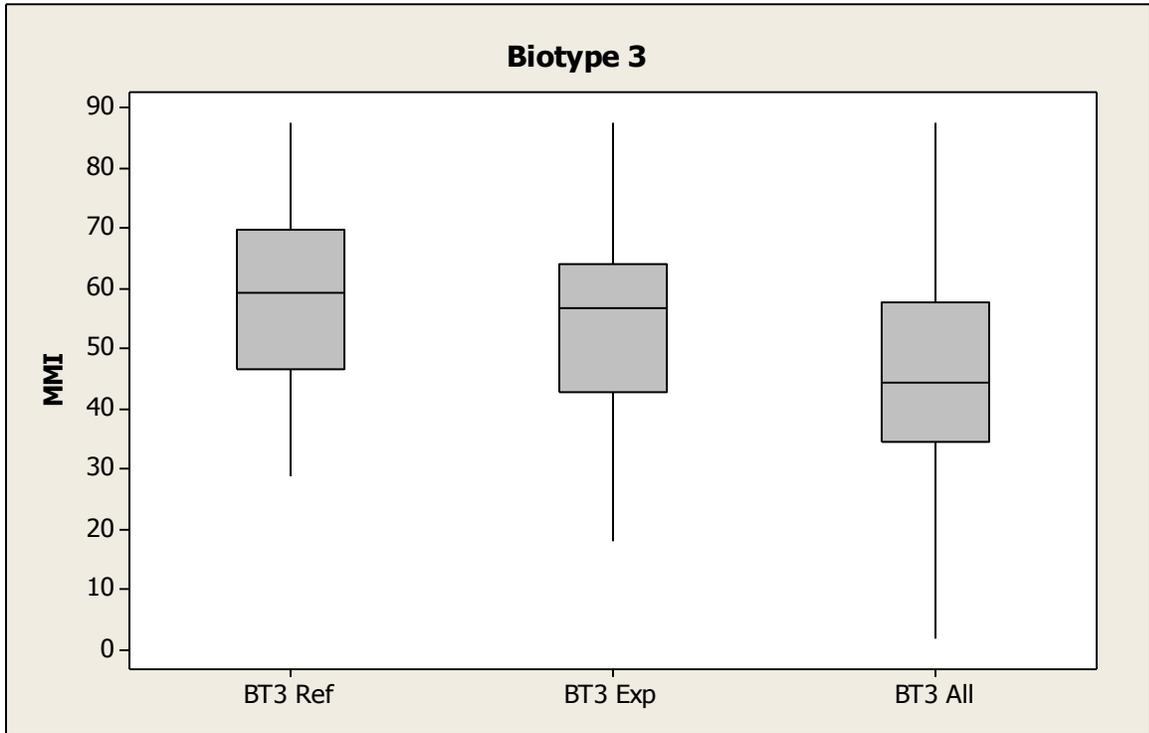


Figure 6. Boxplots of MMI scores for sites in biotype 3 aggregated by group membership (Reference, Expanded, All).

Biological Thresholds Options

Three options are developed; they differ in terms of group membership for use support and definition of normal range. For each option, thresholds are determined and the potential implications for streams statewide are considered. Option 1 defines use support based on a narrow operating range (central 90%) of MMI values at reference sites. Option 2 defines use support based on a normal operating range (central 95%) of MMI values at reference sites. Option 3 defines use support based on a normal operating range (central 95%) of MMI values at sites in the expanded set. Group membership for option 3 in biotype 1 was modified to reduce the number of sites by making the human disturbance criteria slightly more restrictive than values recorded for the group of reference sites. The three options define a range of policy options for characterizing use support (Table 2). In addition, it is clear that large increases in sample size can diminish the width of the gray zone.

Biotype	Option	N of Sites	Average	Standard Deviation	Attainment Threshold	Impairment Threshold
1	1	42	68.37	10.60	54.6	45.9
1	2	42	68.37	10.60	51.6	42.0
1	3	174	59.28	18.28	27.1	19.1
2	1	61	69.70	11.77	53.8	45.9
2	2	61	69.70	11.77	50.4	41.6
2	3	108	67.11	12.81	45.2	38.1
3	1	29	58.30	13.98	41.0	27.0
3	2	29	58.30	13.98	37.1	21.6
3	3	57	54.64	14.26	31.4	20.4

Table 2. Summary of group characteristics and biological thresholds thresholds for each option and each biotype.

The Division recommends adoption of criteria derived from Option 2. From a regulatory perspective, it is based on a methodology that most closely matches EPA guidance; from a practical perspective, it yields thresholds consistent with other evidence of use support (see next section). In subsequent sections of this document, all attention is focused on the Option 2 criteria.

Perspectives on Use Attainment

One of the challenges inherent in developing biological thresholds is that there are few direct measures of use support. Inferring use support from biological condition at reference sites is a reasonable starting point, but it may result in a more exclusive set of sites than is necessary. Too much exclusivity in group membership is likely to underestimate variance, which may lead to an unrealistic basis for developing thresholds.

Although there is no *a priori* basis for deciding what constitutes use support for the macroinvertebrate community, use support may be inferred from data available for some fish communities. For example, the Division of Wildlife applies special designations to stream segments that support native trout populations, naturally-reproducing trout populations, and gold

medal fisheries. Self-sustaining wild trout populations, especially for the native trout species, require good quality habitat and an adequate food base (i.e., bugs). The gold medal fisheries, which are highly prized by anglers, represent “the highest quality cold water habitats”⁸ that DOW seeks to protect. The Division believes that these special fisheries designations are direct evidence for support of the aquatic life use.

For many of the segments with special fisheries designations, biological condition also can be determined from existing macroinvertebrate data. The set of MMI values for Gold Medal fisheries in Biotype 1, for example, can be compared with other use support information from other sources (e.g., the set of reference sites). MMI values in streams with special fishery designations are generally lower than those for reference sites in the same biotype (Figure 4- Figure 5). The differences are not large, but it would be hard to argue that the aquatic life use is not being supported in these special fisheries where biological condition is apparently below that of reference.

Evaluation in the “Gray Zone”

As explained previously, using interval and equivalence tests to derive biological thresholds thresholds creates a gray zone within which neither hypothesis is rejected. Although the statistical evaluation does not lead to a clear-cut answer, MMI values are low enough that more information should be considered before reaching a conclusion about impairment. A key piece of information is the aquatic life class for the segment in question.

Within each of the two major aquatic life use classifications – cold water and warm water aquatic life – streams may be designated as class 1 or class 2 based on expectations for the abundance and diversity of the biota. Class 1 streams support “a wide variety of cold water biota, including sensitive species” whereas class 2 streams do not. The presumption is that a “substantial impairment of the abundance and diversity of species” occurs as the result of “physical habitat, water flows or levels, or uncorrectable water quality conditions.”

When the MMI score falls within the gray zone and it is from a class 2 stream, the stream is considered to be in support of the use because expectations are diminished. However, the same score in a class 1 stream would elicit a different response requiring examination of additional data. The Division proposes evaluation of two auxiliary metrics that were not included in the MMI. These two metrics – Shannon diversity and the HBI – were selected because they provide information on the two characteristics – diversity and sensitivity – that separate class 1 and class 2 streams.

Evaluation of the auxiliary metrics is based on the equivalence test, rather than the interval test or both, in order to require a convincing demonstration that, in spite of the low MMI, the site is not in bad shape. If the null hypothesis of the equivalence test is rejected, it means that the auxiliary metric score is equivalent to those of the reference distribution. In addition, because the regulation addresses diversity *and* sensitive species, a site will be considered impaired when

⁸ Colorado Wildlife Commission Policy on Wild and Gold Medal Trout Management; September 18, 1992, revised Jun 12, 2008.

the MMI is in the gray zone and either auxiliary metric fails to show equivalence to the reference set.

An example is given for evaluation of attainment for each of the auxiliary metrics. The example is derived from one of the three biological thresholds options described above, but the logic is transferable to any of the options.

Shannon Diversity Index

The diversity of taxa observed at each site is captured with the Shannon index, which is calculated in EDAS for every sample. For each biotype, diversity values were assembled for all reference sites and the distributions were examined (Figure 7). After removing one outlier from the lower end of the distributions for biotypes 2 and 3, the three distributions were found to be normal and the variances were not significantly different. Means were significantly different, however, meaning that equivalence thresholds must be established separately for each biotype.

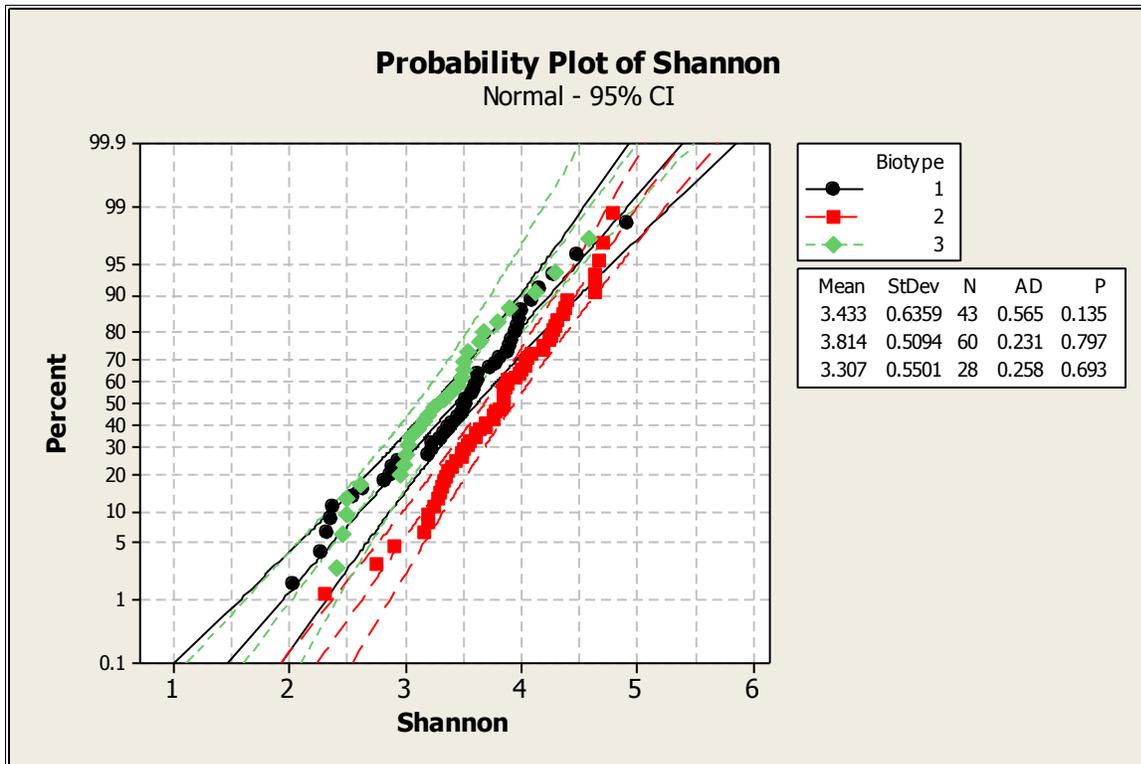


Figure 7. Probability plots for Shannon diversity values at reference sites in each biotype.

Using option 2 (as described above), the dimensions of the gray zone and the equivalence threshold are presented for each biotype in Table 3. If a site in biotype 2 had an MMI score of 45, for example, it would fall in the gray zone and be considered possibly impaired. A diversity score 2.8 would lead immediately to a conclusion of impairment. A diversity score greater than 3.0 would be a necessary, but not a sufficient, condition for considering the site in attainment. A decision regarding attainment could not be reached, however, until the HBI also had been evaluated.

Biotype	Attainment	Impairment	Shannon	HBI
1	51.6	42.0	2.4	5.4
2	50.4	41.6	3.0	5.1
3	37.1	21.6	2.5	7.7

Table 3. Boundaries of the gray zone and thresholds of equivalence for two auxiliary metrics – Shannon diversity and HBI.

Hilsenhoff Biotic Index

The Hilsenhoff Biotic Index (HBI) is a widely used indicator of organic pollution. High values of the index indicate a predominance of tolerant organisms (i.e., the sensitive species have been lost). Like the Shannon index, the HBI is not part of the MMI and the values are routine output from EDAS. Distributions of the HBI are normal for reference sites in each of the biotypes and there are no outliers (Figure 8). Variances were not significantly different among the biotypes, but means were. Accordingly, equivalence thresholds are developed separately for each biotype (Table 3).

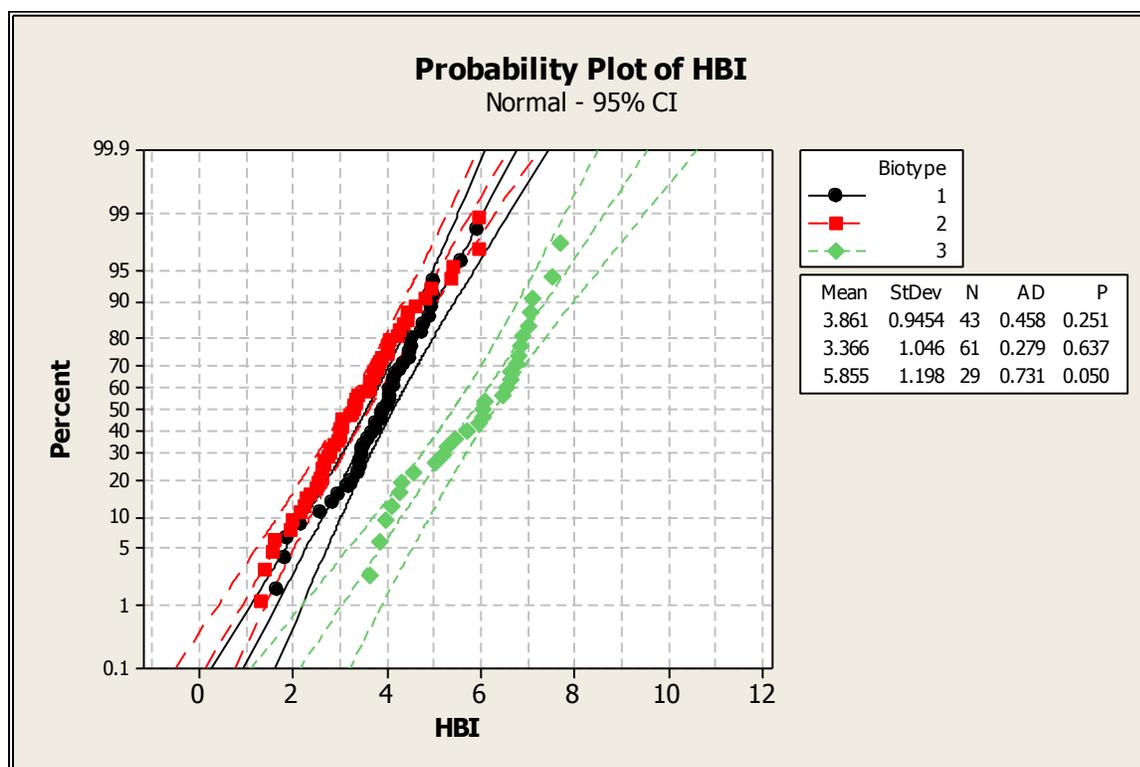


Figure 8. Probability plots for HBI values at reference sites in each biotype.

In contrast to the diversity index, low scores of the HBI indicate better conditions. Consequently, using the same example presented in the preceding section, the HBI would have to be *less than* 5.1 for the site to be considered in attainment. Moreover, both auxiliary metrics would have to show attainment before the site, which had an MMI in the gray zone, would be

considered in attainment. If either auxiliary metric failed to show attainment – if the HBI was at least 5.1 *or* the Shannon index was not greater than 3.0 – the site would be considered impaired.

Protection of High-Scoring Sites

The MMI thresholds that have been presented as options for biological thresholds are appropriate for detecting impairment on the basis of individual samples. However, because some sites have much higher MMI scores, it would seem to leave open the possibility of undetected degradation at high-scoring sites. For example, a site with an MMI value of 90 in biotype 2 would have to experience a drop of more than 40 points before it would be declared impaired. It would be prudent to take action sooner if significant degradation was occurring.

In order to detect a significant decline in biological condition, there must first be an understanding of what constitutes a normal range or year-to-year variability in MMI scores at sites where there is no known water quality or habitat trend. Fortunately, there are a number of sites where biological condition has been examined more than once over a period of years. The change in MMI (d-MMI) was calculated for all available pairs, after excluding any data from the drought year (2002). Because the typical time window for assessments is usually five years, and because a short interval diminishes the likelihood that the change will be influenced by a trend, the data set was restricted to pairs spanning five or fewer years. Statistical comparison of d-MMI values for different intervals (1, 2, 3, 4, or 5 years) showed no significant difference; therefore, all pairs were combined for analysis. For sites with more than two samples, no sample was used more than once, and preference was given to shorter intervals (but at least a year).

Distributions were examined first for each biotype. Distributions were normal, and two outliers were removed from biotype 1. After outliers were removed, variances were found to be homogeneous and there was no significant difference among the biotypes. Accordingly, d-MMI scores for all three biotypes were combined (mean=-0.20; s.d.=12.93; N=123). The d-MMI scores show considerable variability. The “normal operating range” is 25.3 to -25.3⁹. For a change to be considered significant at the 0.05 level, the MMI would have to increase or decrease by 25.3 points. The inter-year variability may seem relatively large, but the standard deviation is similar in magnitude to those observed among reference sites in each of the three biotypes.

⁹ The normal operating range is centered on mean of zero to reflect the assumption that no change in biological condition would yield no change in MMI. The observed mean of -0.20 MMI units is indistinguishable from zero.

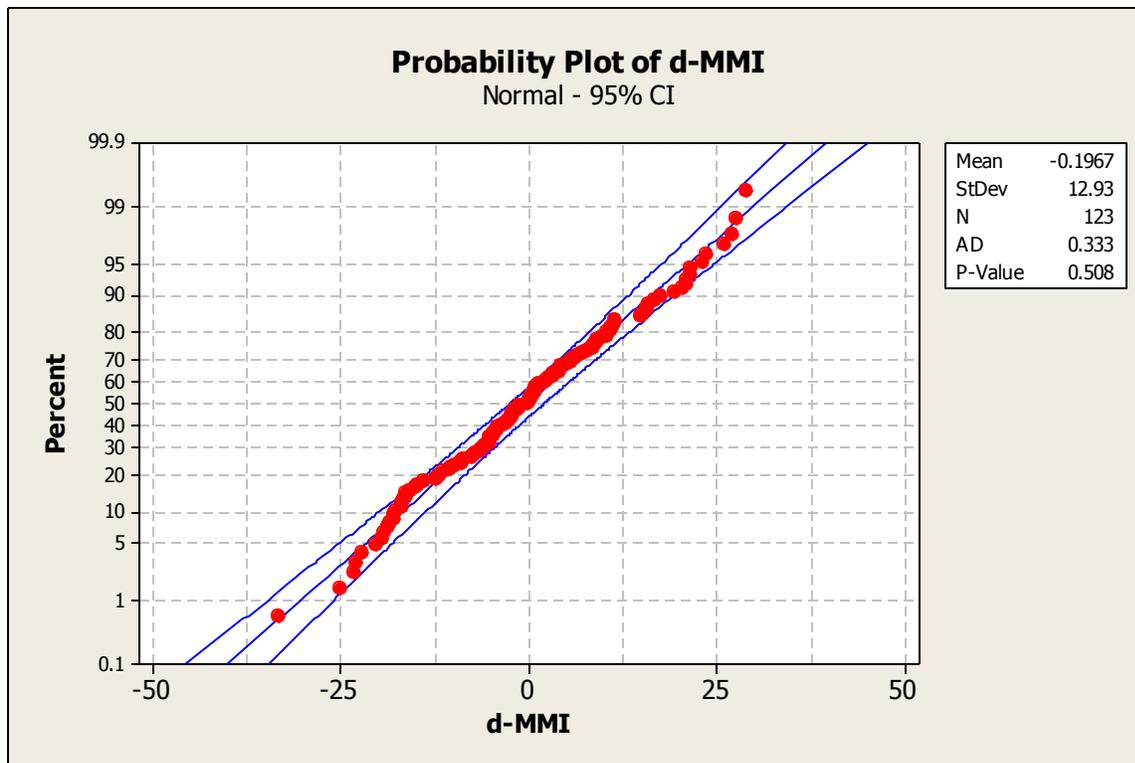


Figure 9. Probability plot of d-MMI scores after removing two outliers as described in the text.

The threshold for decline in MMI can be derived from the same interval and equivalence tests applied previously to set thresholds for each of the biotypes. Using the normal operating range (central 95% of the d-MMI scores) and a 5% level of significance for the interval and equivalence tests, the thresholds for attainment and impairment would be set at -22.3 and -29.0 MMI units, respectively. Given the required magnitude of the change and the high initial level for biological condition, the Division proposes a binary decision system based solely on the equivalence test. If the MMI declines by at least 22.3 units, it should register a serious concern about biological condition, and the site should be considered impaired.

Application of this test proposed for protection of high-scoring sites use requires having at least two MMI scores. Although the test value was derived from samples separated in time by one to five years, there is no need to apply similar restrictions for the regulatory decisions. As a practical matter, the initial MMI must be at least 22 MMI units higher than the impairment threshold for the appropriate biotype (see Table 2). Thus, the initial MMI would have to exceed 64 for biotypes 1 and 2, or 44 for biotype 3, before a site would qualify for consideration as high-scoring.