I. Overview of the Water Quality Assessment Process

A. Water Quality Policy 1-11 (revised September 2002)

ii. Additional Clarification of the Binomial Distribution Method

Background
The 1998 303(d) list was based on Policy 1-11 (revised June 1997), which treated all conventional pollutants similarly for assessment purposes. The 1997 policy stated that data received for assessment must meet the following criteria:

> "For water measurements of temperature, dissolved oxygen, pH, turbidity and total dissolved oxygen in 10% or more of the measurements and a minimum of at least two measurements are beyond the numeric state surface water quality criteria within the most recent 5 year period that the data has been collected."

The 1997 policy decision was based on EPA guidelines suggesting that a stream segment be listed as impaired when greater than 10% of the measurements of water quality conditions exceeded numeric criteria. This is sometimes referred to as a "raw score" assessment method.

During and after the 1998 listing process, Ecology received comment from communities and regulated industry that the 303(d) listing criteria, using the raw score assessment approach, erred significantly on the side of creating 'false positive" listings (that is, listing a water body as impaired when in fact uses are being attained). These concerns were heightened by the regulatory restrictions placed on point source NPDES permit holders if they discharged into a 303(d)-listed water body that was listed for a parameter included in their discharge (see Water Quality Program Permit Writers Manual, Chapter 6 at pg.30). Therefore the significance of falsely listing a water body as impaired leads to increased regulatory burdens and costs for permit compliance. These concerns may also have been heightened by past listing decisions in which a water body was listed based on sparse raw data, but then EPA requires a higher bar for "de-listing" a water body, to prove that the water is, in fact, not impaired.

Based on the feedback that Ecology appeared to be unfairly listing waterbodies as impaired, at an unreasonable cost to communities and regulated industry, Ecology decided to examine more closely what data we use for 303(d) listing and how listings are determined. The goals Ecology wanted to achieve in this examination were to:

- Reduce the error rate of 303(d) listing decisions
- Work to better understand variations in datasets and how this information should affect the prioritization and scope of subsequent TMDLs.
- Write policy to guide an accurate assessment of Washington waters even when small datasets are all that are available.

Assessment of water quality conditions has a certain degree of uncertainty associated with it. Due to the variability introduced by these conditions, an assessment decision may possibly being wrong by way of false positive and false negative errors. A false positive error occurs when a stream is listed, in spite of the fact that it meets the water quality standards and thus is protecting beneficial uses. A false negative error occurs when a stream that is not meeting water quality standards is listed.

The "Raw Score" vs. "Binomial Distribution" Methodology

To achieve a higher confidence level in 303(d) listing decisions, Ecology researched various methodologies for 303(d) listing decisions and their associated confidence levels for ensuring an accurate assessment. In particular, Ecology reviewed information available for using the binomial distribution model. Ecology considered percent violations, confidence levels, and minimum number of samples, samples, and the required number of exceedances of this method to determine the applicability to Washington's 303(d) listing process.

Assessment of water quality conditions has a certain degree of uncertainty associated with it. Due to the variability introduced by these conditions, an assessment decision may possibly being wrong by way of false positive and false negative errors. A false positive error is when a stream is listed, but meets the water quality standards and thus is protecting beneficial uses. A false negative error is when a stream is not listed, but in fact does not meet the water quality standards.

Washington's 1997 listing policy used the raw score method described in the section above. This method does not set minimum sample sizes. The raw score method does not consider the likelihood and costs of making an erroneous listing (a false positive) or not listing a water when it is truly impaired (false negative).

The assessment challenge is to determine, with limited amount of sample data, whether an apparent violation of standards warrants listing a segment as impaired. Likewise, limited data must be relied upon to determine whether actions taken to address water quality degradation have the desired results. To overcome these challenges and decide what method to use for adequately assessing persistent impairment of a water body, several facts must be taken into account.

- Water quality samples are taken from a variety of water quality conditions.
- Concentrations of pollutants can vary naturally,
- Errors can be made in measurements.
- Occasional violation of a numeric standard may not be an indication that uses are not being attained.

These are especially true for conventional pollutants (such as temperature, dissolved oxygen, and pH) that, at times, will naturally exceed the criteria based on natural hydrologic, geographic, and climatic conditions. An occasional violation, even from an anthropogenic source, may not be detrimental to uses of an aquatic environment. Therefore, when considering what assessment method to use, it is reasonable to take a statistical approach to account for the probability of errors.

Ecology determined that a better approach would be to set data quality objectives for these error rates and derive a minimum sample size. This approach is often termed the "binomial distribution" method. The use of the binomial approach greatly reduces the error of listing false positives, as shown in a paper published in the *Environmental Science and Technology* (Smith et al, Feb. 2001).

Relative to the raw score method, the binomial approach is more prone to false negative errors. This tendency towards false negative errors can be mitigated by increase sample sizes, although it is still higher than the raw score approach. Ecology considered this difference when weighing out the advantages of the binomial approach over the raw score method, and determined that the risk of having greatly increased probability of false positives with the raw score method outweighed the risk of increased false negative errors. This decision is supported by the recognition that addressing an identified water quality problem for a 303(d) listed water is a complicated and potentially expense endeavor. TMDL planning can be costly, and the followup implementation of a TMDL can also impose additional costs. Given the limited state resources available, and the importance of focusing on the state's highest priority problems, Ecology believes it is important that waters truly impaired be identified.

The use of the binomial approach also allows us to make credible determinations that a water body is impaired even if the sample size is small, thus also reducing the likelihood of not listing false negatives. The binomial method is also non-parametric, which allows the same statistical criteria to be applied to all water quality constituents without an estimate of variance.

When a binomial approach is used, error rates can be explicitly managed in the assessment by considering the number of samples taken, determining the acceptable and unacceptable violation rates and selecting the cutoff values for declaration of impairment. Ecology recognizes that these decision should be governed by the properties of the pollutant, the uses of the water segment, and the consequences of false positive and false negative results in the assessment.

For a given pollutant in a water body, the sample proportion of exceedances is a point estimator of the true exceedance probability for the pollutant. Since the estimator varies in a random manner from sample to sample, inferences about the true exceedance probability based on the estimator will be subjected to uncertainty. The degree of uncertainty depends on the exceedances and the sample size: the smaller the sample size is, the greater the uncertainty will be. Therefore, the sample proportions of exceedances should not be used for the determination of water body health without considering its sample size.

Sample sizes

The raw score approach often relies on limited, binary information to make an impairment determination. The binomial approach, however, focuses on the probability of a violation, using the same information. Relative to the raw score approach, the binomial method is more prone to false negatives and less prone to false positives. The tendency towards false negative errors in either approach is mitigated by increasing the minimum sample size. The advantage of the binomial approach is that in addition to controlling false negatives through sample size, false positive errors can be controlled by selecting a violation cut-off.

One of the biggest challenges faced by Washington, and likely many other states, is that listing decisions must be made on available data, which for many waterbodies can be quite small. Given these typically small sample sizes, any proposed listing and delisting methods, based on the calculated sample exceedance percentages, must be applicable to both large and small samples. When estimating the true exceedance probability of a pollutant or testing hypotheses about the true exceedance probability, it is suggested that 10 or more sample measurements (minimum sample size = 10) be required for assessing whether or not a water body reach is impaired based on criterion exceedances, (Smith et al, 2001).

Ecology considered not only the question of how many exceedances are required, but also whether there should be a minimum sample size that we will assess. For example, if a minimum of three exceedances and a minimum of ten samples are required, what do we do if there are four exceedances out of only nine samples? The 1998 list did not have a minimum number of samples, with the exception of fecal coliform. After discussion, Ecology decided that for toxics and all other pollutants except fecal coliform, the policy should not set a minimum number of samples. When the data shows enough exceedances to list, it makes no sense to disqualify that listing merely because there are too few samples that ***do not*** show an exceedance. By default, the minimum number of samples would become the same as the minimum number of exceedances. For toxics, this is a fixed number. For other pollutants, the binomial distribution method with a ten percent confidence level requires a minimum of three exceedances regardless of how few samples are available. Thus, for example, if there were less than ten samples (even if there were only 3), at least three of them would need to show an exceedance for a water body to be listed. This approach provides a higher level of confidence in the assessment when small datasets are all that are available to determine the condition of a water body.

Confidence Levels/Error Rates

The concept of the binomial distribution methodology is to statistically determine the probability that the water body as a whole, rather than a percent of the samples, are impaired. Using 10% as an example, the goal is to determine ten perncent exceedances of a water quality standard in the water body as a whole, rather than ten percent exceedances in the samples taken from that water body. The samples are the best evidence regarding the condition of water body, but are not the entire water body. Thus, a “safety margin” (based on a statistical standard deviation) is required in order to declare, with a given degree of confidence (defined by the confidence interval), that a set of randomly collected samples which meet the criterion accurately show that the water body itself meets the criterion.

Using the binomial approach, one is able to test the hypothesis that the probability of not exceeding the standard is greater than or equal to 90% and the probability of exceeding the standards is less than ten percent. One must also define the minimum detectable difference. Ecology uses the EPA percentiles for determining beneficial-use support, (e.g. over 90% is fully supporting uses)

When sample sizes are around 20-25, the assessment process can confidently rely on statistical procedures to manage and measure both false positive and false negative errors. It has also been recognized that false negative errors are more likely to occur with the statistical methods than with the raw score approach. While increased sample size will reduce the probability of false negatives, one strategy for reducing the error rate in false negatives would be to increase the error rate for false positives.

However, the raw score method does not explicitly manage error rates. The raw score is conceptually similar to the binomial approach. Both methods use the number of violations as the test statistics. However, the raw score is a poorly designed test statistic. The raw score approach results in an unusually large error rate in false positive listings, regardless of sample size. As sample sizes increase, the errors for false negatives is reduced but the average error rate is still large in the raw score method.

If the binomial distribution method is used, a decision must be made concerning the required confidence level. Other states have used between 80%-95%. A 95% confidence level is commonly used for high quality statistical reports.

A higher confidence level leads to fewer mistakes of listing false positives, but in turn can lead to false negatives, or failing to list waters that need attention. A lower confidence level will lead to more impairment listings, including more false positives, but also fewer failures to list when the water should be listed.

The following tables show the minimum sample sizes and the number of exceedances to give a 90% and 95% confidence level that the water will be impaired. For comparison purposes the raw data approach used in the1998 303(d) list (straight 10% exceedance of samples with a minimum of 2) is also included.

90% Confidence

Sample Size	Number of Exceedances to List
0-2	NA
3-11	3
12-18	3
19-25	4
26-32	5
33-40	6
41-47	7
48-55	8
56-63	9
64-71	10
72-79	11
80-88	12
89-96	13
97-100	14

95% Confidence

Sample Size	Number of Exceedances to List
0-3	NA
4-14	4
15-20	5
21-27	6
28-34	7
35-41	8
42-48	9
49-56	10
57-63	11
64-71	12
72-79	13
80-88	14
89-96	15
97-100	16

Straight 10%

Sample Size	Number of Exceedances to List
0-1	NA
2-20	2
21-30	3
31-40	4
41-50	5
51-60	6
61-70	7
71-80	8
81-90	9
91-100	10

Decisions on Using the Binomial Distribution Methodology

After studying the statistical approaches to more accurately determine impaired water and compare them to the raw data method, the draft Policy 1-11 included the binomial distribution method for the 2002/2004 listing process for pH, turbidity, total dissolved gas, phosphorus, nitrogen, and hardness (and any other pollutant besides temperature, dissolved oxygen, toxics, and fecal coliform),. This slightly raises the number of exceedances required for listing, compared to the 1997 listing policy.

The concept is to apply the 10% standard in the 1997 listing policy to the *water*, not the sample data, by using a binomial distribution statistical method. Waters would be placed on the 303(d) list if enough of the data violate the standard so that there is a 90% confidence level that the water body from which the samples were taken violates the standard 10% of the time. The effect is to increase the *percentage* of the samples that are required to be violations – only a slight increase when there are many samples, and a larger one when there are very few samples.

During deliberations on using this statistical approach, Ecology recognized that larger sample sizes give a greater probability that a water body is truly impaired and helps to prevent false positive listings, while the raw data method has a significantly higher probability of a false positive listings. However, we also recognized that requiring larger sample sizes before determining impairment would not allow the use all available information. Therefore, we made a decision to allow smaller sample sizes, despite the recognition that it could lead to a much higher degree of false positives and false negatives. To mitigate this risk we decided to use the binomial distribution method to ensure a higher degree of certainty that the water is indeed impaired.

The goal of the binomial approach, as well as the straight 10% approach it replaces, is to define persistent pollution that is expected to impair beneficial uses, rather than basing listing on a single sample or on a very short period that violates the water quality standards. The 1997 listing policy put waterbodies on the 303(d) list when 10% of the samples exceeded the water quality standards. In the 2002 policy, the binomial distribution instead tests the hypothesis that the actual conditions in the water body are such that at least 10% of the water would show an exceedance of the standard. This slightly raises the number of exceedances required for a 303(d) listing. It also increases the degree of certainty that there is a persistent pollution problem in the water body before listing.

The most significant effect from this change occurs with a small sample size. In this case, only a few more exceedances are required (in fact, with *very* small sample sizes, only one additional exceedance is required), but these few more exceedances produce a much higher *percentage* of exceedances and thus much provide greater confidence that the measurements truly reflect a water quality problem, as opposed to sampling error or random fluctuations. Ecology believes that, when there are fewer data points available, it is appropriate to require a relatively stronger showing of a problem, without raising the bar for listing to an excessive level. For this reason, a minimum number of three exceedances is required for putting a waterbody segment on the 303(d) list.

During the deliberations on revisions to the draft listing policy, discussions were held on using the binomial distribution method for removing waterbody segments from the 303(d) list, and whether the process should be defined differently. In general, agreement was reached that a water body that was previously listed would not be listed for the 2002/2004 list if:

- The standards or criteria have changed so that the original data would not now lead to listing;
- Further information about thedata quality calls the listing into question.
- After closer analysis,the previous listing decision is shown to have been made in error.

Ecology chose to follow the EPA guidelines set forth for the 2002 assessment by adopting the integrated report and creating 5 assessment categories. In this endeavor, Ecology recognized the importance of consistency in regard to current and previous listings. Therefore a new 303(d) list was compiled including a reassessment of previous data based on the revised policy. If a water body was previously listed and the data used for that original listing still meets the new listing criteria, the water body was kept on the list until new data show that it should not.. Because the integrated report includes category 2, *Waters of Concern*, assessments that showed some evidence of impairment but not enough to be included on the 303(d) list will not "fall off the radar screen".Category 2 identifies those waters which require further study todetermine the actual status of impairment to a more certain degree. This allows the full results of the Water Quality Assessment categories, and the 303(d) list, to be used as an appropriate management tool for identifying those water bodies, and only those water bodies, where it is strongly shown that TMDLs are needed.

Public Comments on the Proposed Binomial Distribution Methodology

Public comments were received on revisions to Policy 1-11 in July 2002 that included the binomial distribution method for total dissolved gas, pH, nitrogen, phosphorus, turbidity, and hardness. In general very favorable comments were received from the public on improving upon the methodology to list based on the representativeness of the water body as opposed to the number of samples taken. EPA expressed appreciation of the state's efforts to improve the representativeness of sampling in waterbody segments and desire to increase confidence about representing the condition of the water body as a whole.

However, EPA also expressed concerns that the binomial approach might result in under-reporting impaired waters. Overarching was EPA's concern that Ecology's adoption of the proposed binomial approach without revising their data collection strategy would result in many waters being excluded from listing on the 303(d) list due to insufficient data. EPA was also concerned that the proposed binomial approach requires a high percentage of samples to exceed the criteria before impairment is confirmed, higher than the 10% threshold suggested in EPA's 305(b) guidance. EPA feels that the binomial distribution method should be used only when more than 500 data points are available, where the sample exceedance approaches 10%.

Ecology's Response

The goal of the binomial approach, as well as the straight 10% approach it replaces, is to define persistent pollution that is expected to impair beneficial uses, rather than basing listing on a single sample or on a very short period that violates the water quality standards. The 1998 policy put water bodies on the 303(d) list when 10% of the samples exceeded the water quality standards. In the 2002 policy, the binomial distribution instead tests the hypothesis that the actual conditions in the water body are such that at least 10% of the water would show an exceedance of the standard. This slightly raises the number of exceedances required for a 303(d) listing. It also increases the degree of certainty that there is a persistent pollution problem in the water body before listing.

The most significant effect from this change occurs with a small sample size. In this case, only a few more exceedances are required (in fact, with *very* small sample sizes, only one additional exceedance is required), but these few more exceedances produce a much higher *percentage* of exceedances and thus much provide greater confidence that the measurements truly reflect a water quality problem, as opposed to sampling error or random fluctuations. Ecology believes that, when there are fewer data points available, it is appropriate to require a relatively stronger showing of a problem, without raising the bar for listing to an excessive level. Given the level of scrutiny that the 303(d) List undergoes, and the recent legislative efforts in the state (such as the Credible Data Bill) to make listing decisions more difficult to list waters at all, we believe it is critical to ensure that decisions we make to list waters based on small sample sizes are credible. The binomial method gives us this assurance.

<u>Conclusions (excerpt from Policy 1-11, revised September 2002)</u>

Other Pollutants

For total dissolved gas, pH, nitrogen, phosphorus, turbidity, and hardness (and any other pollutants besides toxics, fecal coliform, temperature, and dissolved oxygen), the assessment decision is based on persistence of the pollutant at levels in excess of the water quality standard. The criterion for persistence is when an exceedance of the standard is indicated for ten percent of the water in the segment. This can be understood as addressing all of the water samples that theoretically could be taken from the segment, as opposed to only the water samples actually taken. The test is whether, with a given degree of confidence, the set of randomly collected samples accurately show that the water that the samples were taken from has a true exceedance percentage of at least ten percent.

The true exceedance percentage will be determined using a binomial distribution method with a 90 percent confidence interval. A segment will be placed on the 303(d) list if, in applying this method, the data show a true exceedance percentage in the waterbody segment of ten percent or greater. This method requires somewhat more than ten percent exceedance of standards of the water samples. The precise number of exceedances required depends on the sample size. With a smaller sample size, a higher percentage of the samples must be exceedances to support a listing. With a larger sample size, the percentage of exceedances required to support a listing is lower and approaches ten percent. Table 1 gives the exact number of exceedances required for sample sizes of up to 500 samples. With very small sample sizes, a minimum of three exceedances is required.

A segment will be place in the *Waters of Concern* category if the number of exceedances is below the minimum required to place it on the 303(d) list, but is five percent or more of the samples.

Table 2.[1] Minimum number of exceedances required to place a waterbody segment on the 303(d) list, using a binomial distribution, with a 90% confidence that the true exceedance percentage in the waterbody segment is greater than or equal to 10%, for 1-500 samples.

Sample Size	Minimum Number of Exceedances	Sample Size	Minimum Number of Exceedances	Sample Size	Minimum Number of Exceedances
1-2	NA	157-164	21	334-343	41
3-11	3	165-173	22	344-352	42
12-18	3	174-182	23	353-361	43
19-25	4	183-191	24	362-370	44
26-32	5	192-199	25	371-379	45
33-40	6	200-208	26	380-388	46
41-47	7	209-217	27	389-397	47
48-55	8	218-226	28	398-406	48
56-63	9	227-235	29	407-415	49
64-71	10	236-244	30	416-424	50
72-79	11	245-253	31	425-434	51
80-88	12	254-262	32	435-443	52
89-96	13	263-270	33	444-452	53
97-104	14	271-279	34	453-461	54
105-113	15	280-288	35	462-470	55
114-121	16	289-297	36	471-479	56
122-130	17	298-306	37	480-489	57
131-138	18	307-315	38	490-498	58
139-147	19	316-324	39	499-500	59
148-156	20	325-333	40		

References

Smith, E., Ye, K., Huges, C., and Shabman, L. February 2001. Statistical Assessment of Violations of Water Quality Standards Under Section 303(d) of the Clean Water Act. Department of Statistics and Department of Agricultural Economics, Virginia Tech.

Lin, P., Meeter, D., and Niu, X. October 2000. A Nonparametric Procedure for Listing and Delisting Impaired Waters Based on Criterion Exceedances. Department of Statistics, Florida State University.

[1] An error was found in Table 2 of Policy 1-11, revised September 2002. The table was corrected on 4/16/05 and all affected pollutant parameters were reassessed based on the corrected table.