White Paper

Integrated Reporting Improvements

Statistical Methods for Listing and Assessment of Large and Long Term Data Sets

To: Stakeholder Work Group
From: DEQ IR Improvement Project Team

Author Names: Becky Anthony, Basin Specialist, Western Region James McConaghie, PhD., Water Quality Standards and Assessments

Revision Date: Friday, Nov. 20, 2017

Water Quality Standards and Assessments
700 NE Multnomah St.
Suite 600
Portland, OR 97232
Phone: 503-229-5696
800-452-4011
Fax: 503-229-5850
Contact: Jennifer Wigal

www.oregon.gov/DEQ

DEQ is a leader in restoring, maintaining and enhancing the quality of Oregon's air, land and water.

Summary

The Oregon Department of Environmental Quality currently assesses compliance with numeric biologically-based aquatic life criteria for toxic substances using a simple two sample threshold for impairment. If any two samples within a waterbody exceed the criteria value, it is grounds for listing a waterbody as impaired (Category 5) on the 303(d) list. Acute criteria are expressed as a 1-hour average concentration, and chronic criteria are expressed as a 4-day average concentration, even though routine monitoring samples are collected bimonthly and there is rarely enough information to confirm what the duration of exposure is for any individual sample.

Under this method, especially for the more stringent chronic criteria, the chance that the concentration of any two samples will be higher than the criterion magnitude increases solely as a function of sample size. If a determination is made that the waterbody is not supporting a designated beneficial use as a result of an exceedance the waterbody is listed as impaired on the 303(d) list and identified as needing a TMDL as well as other regulatory actions that are triggered for the waterbody.

DEQ has received significant input from a number of stakeholders that the current assessment method for toxic substances overestimates the number of impaired waters, particularly in circumstances where waterbodies are listed based on only two samples exceeding the criterion magnitude when there is also a large number of attaining samples. The validity of listings has been questioned and concerns raised regarding potentially significant resource burdens to address impairments that may be in error. Additionally, the policy is known to have disincentivized the submission of long-term data sets by stakeholders to DEQ's call for data.

More reliable statistical methods for evaluating attainment of water quality standards are available, approved by EPA, and have been implemented in other states. For relevant data sets, DEQ is proposing to use a binomial hypothesis test that accounts for sample size, errors in sample accuracy and precision, and explicitly defines the acceptable levels of certainty in making a determination. Using this method, the risk of making errors in determining both impairment and attainment is defined and can be weighed.

DEQ is proposing to update the listing and delisting methodology for numeric water quality standards for toxic substances and conventional pollutants. Revisions to the methodology for assessment of criteria where a statistical threshold or proportion used to determine attainment is already clearly expressed in Oregon's water quality standards are not being considered at this time. The standards include temperature, bacteria, and continuously monitored dissolved oxygen. The only standards considered here are numeric criteria for toxic substances (OAR 340-041-8033, Table 30), conventional pollutants (pH, instantaneously measured dissolved oxygen), and human health toxics criteria (OAR 340-041-8033, Table 40).

Background

Components of Water Quality Standards

Water quality standards have three components: designated beneficial uses, water quality criteria to protect the uses, and anti-degradation policies.

Numeric water quality criteria also have three components: magnitude, duration, and frequency. A water quality standard is considered exceeded if the average concentration of a waterbody is greater than the allowable magnitude longer than one hour, for acute criteria, or 96 hours, (four days) for chronic criteria, more often than once every three years on average.

Magnitude

The magnitude of a criterion is the value of the concentration threshold of the pollutant determined to be protective for a specific beneficial use and context. Most water quality criteria have separate thresholds to protect against short term (acute) and long-term (chronic) exposure to pollutants.

Acute criteria are calculated using half of the final acute value, which is usually the concentration that produces 50% mortality in the study species (LC-50) of the 2 most sensitive genera from published toxicity studies[1]. Chronic criteria are based on delayed lethal and sub-lethal effects (changes in mobility, growth, or reproduction) observed in 10% or 20% of the sample population (EC10 / EC20) from long-term exposure studies. If long term studies are not available, the chronic toxicity is derived from the acute toxicity with an empirical acute to chronic ratio (ACR) that is species-specific.
An individual sample that exceeds the criterion magnitude is an excursion[2], but is often also referred to as an exceedance or sample exceedance. A sample exceeding the magnitude is not the same as the waterbody exceeding the standard.

[1] EPA 1994, Water Quality Standards Handbook: Second Edition. Appendix H. Derivation of the 1985 Aquatic Life Criteria. United States Environmental Protection Agency, Office of Water. EPA 823-B-94-005a. August 1994.
[2] EPA 1985, Technical Support Document for Water Quality based Toxics Control. EPA-440/4-85-032, United States Environmental Protection Agency. September 1985.

Duration

The duration component of a water quality criterion is the length of time aquatic life may be exposed to a concentration above the criterion threshold before the criterion is considered exceeded. This component of the criteria accounts for short-term excursions of pollutant concentration where impacts are expected to be minor or where the community can quickly recover. The duration of exposure used for most toxic substances is a 1-hour acute exposure and a 4-day (96-hour) chronic exposure. Acute exposure is based on direct mortality of aquatic life within a short time period. Chronic exposure is based on delayed lethal or non-lethal effects that impair viability, growth, or reproduction of aquatic life. Water samples collected at a high frequency should be averaged to a 1-hour mean or 4-day mean concentration before comparison to the magnitude. A common assumption for instantaneous grab samples collected at intervals greater than the duration periods (e.g. weekly, monthly) is that they are representative of the 1 hour or 4 day mean concentration. The human health criteria are based on long-term average exposure. The duration of exposure is specified in the recommended water quality criteria. Typical exposure durations are lifetime and 30-day.

Frequency

The frequency is the number of excursions that can occur over time without impairing the aquatic community or other relevant designated beneficial use. Most recommended national water quality criteria (304(a) criteria) for toxic substances specify that the standard is not to be exceeded more than once every three years *on average*. This frequency was selected by EPA because it is statistically impossible to project that an excursion will never occur, and to acknowledge that aquatic communities often exhibit resilience to infrequent excursions above the magnitude. The frequency component is intended to allow inconsequential excursions above the magnitude and to account for uncertainty in the accuracy and representativeness of random samples collected from the waterbody.

According to EPA, this 1-in-3-year frequency is a return interval intended to provide a level of protection equivalent to a 7Q10 design flow condition[3]. The 7Q10 is a common flow statistic for defining low flows as the lowest average instream flow over a period of one week with a recurrence interval of 10 years. There is a 10% probability that there will be a lower flow in any given year.   In another source, EPA points to the 1-in-3-year average frequency of exceedance with the intent of providing for ecological recovery from a variety of severe stresses. The 1-in-3-year frequency is also justified based on observed recovery time of aquatic communities from acute exposures to catastrophic disturbance events such as floods, oil spills, wild fires or pesticide applications.[4,5] The

[3] EPA, 2002. Consolidated Assessment and Listing Methodology (CALM). Toward a Compendium of Best Practices, First Edition. United States Environmental Protection Agency. Section 4-6. July 2002.

[4] EPA 1985, Technical Support Document for Water Quality based Toxics Control. United States Environmental Protection Agency. EPA-440/4-85-032, September 1985.

[5] Niemi, G.J., P. Devore, N. Detenbeck, D. Taylor, A. Lima, J. Pastor, D.J. Yount, R.J. Naiman. 1990. Overview of case studies on recovery of aquatic systems from disturbance. Environmental Management 14(5):571-587.

average recovery time for aquatic populations exposed to toxic substances may be closer to 1 year.[6]

The frequency component of water quality standards does not define a sampling interval in relation to the criterion. This leads to uncertainty in interpreting the frequency component regarding the number or proportion of samples needed to determine whether the waterbody exceeds the criteria more than once in three years on average. States have the latitude to define a critical exceedance rate for samples they will use to evaluate whether the standard is being exceeded.[7] As long as the defined critical exceedance rate is not exceeded more than once in a three year period, the waterbody is determined to be attaining the standard. EPA specifies neither that two samples is the only valid critical exceedance value, nor does it strictly endorse any specific proportion of samples. However, EPA suggests that a 5% rate for toxic substances and a 10% for conventional pollutants would provide the desired level of protection (see Table 2, below).

Interpreting attainment of water quality standards

The national recommended water quality criteria are 'ideal standards'[7].They are described as thresholds never to be exceeded that apply to the condition of every part of a waterbody as a whole. They do not address natural variation or uncertainty in water samples collected from that waterbody. Determining attainment of the waterbody under an ideal standard implies that samples collected through monitoring are capable of depicting the true population of possible samples at all points in the waterbody.

Water quality monitoring programs are not capable of monitoring all points and all variability within a waterbody at all times. Due to budgetary and practical constraints, state monitoring programs collect samples at a frequency often considerably longer than the specified duration of acute and chronic exposure in the water quality standards (e.g. monthly, quarterly, annually etc.). The expectation is that individual observations of water quality (the sample) be extrapolated to represent conditions in the waterbody as a whole (the population). Sampling involves inherent uncertainty due to potential for bias, measurement error, and sampling error. Some of which is based on natural variability.

According to EPA, sources of error and uncertainty include[7]:

- sampling variation and bias due to monitoring design
- temporal and spatial variability in the waterbody
- natural variation among samples
- measurement error of samples
- analytical error or contamination

[6] Gergs et al. 2016. Ecological Recovery Potential of Freshwater Organisms: Consequences for Environmental Risk Assessment of Chemicals. Reviews in Environmental Contamination and Toxicology. 2016 (94) 236: 259-294.

[7] EPA, 2002. Consolidated Assessment and Listing Methodology (CALM). Toward a Compendium of Best Practices, First Edition. United States Environmental Protection Agency. July 2002. p.45

These sources of error can be mitigated through selection of monitoring plans and statistical methodologies to reduce error in decision making. Given the amount and variability of data we actually collect for a waterbody, statistical procedures are intended to test:

1. How certain we are that the samples collected represent average conditions in the waterbody?
2. How certain we are that the samples indicate whether the waterbody as a whole is attaining or exceeding the water quality standard?

The decision to list or de-list a waterbody should ideally be based on the most accurate, representative, and verifiable information possible. The chance of falsely concluding a waterbody exceeds the standard, when it is actually attaining (a false-positive, or Type-I error), or falsely concluding it is attaining the standard when in reality it is exceeding (false-negative, or Type-II error) should always be considered. Where possible, affirmative steps should be taken to account for errors in monitoring design and the selection of assessment methodologies. In the absence of complete data for a waterbody, use of statistics can increase the confidence that accurate decisions are being made and that the conclusions regarding impairment or attainment are defensible.

Common methods for determining attainment of WQ standards

The “>1-in-3-year” critical exceedance frequency

Literal interpretation of the exceedance frequency component of the water quality standard uses two samples as the critical exceedance rate. If two or more sample concentrations are observed over the criterion magnitude in a three-year period, the waterbody is considered to exceed the standard.

The ≥2 sample critical exceedance rate assumes:

- each sample evaluated is representative of the condition of the entire waterbody.
- an instantaneous grab sample accurately represents the 1-hr average concentration for acute criteria.
- an instantaneous grab sample accurately represents the 4-day average concentration for chronic criteria.
- there are no errors in sample measurement or reporting.

The rationale for the >1-in-3-year frequency is based on the historic trend of having limited data and very small sample sizes. When there are relatively few samples collected in a waterbody, (e.g. 4 throughout the year in a quarterly monitoring schedule), and there are at least two sample excursions, it likely indicates a water quality issue that is persistent. However, when a waterbody is sampled more intensively, and there are only a small number of sample excursions, it brings into question whether any excursion is representative of waterbody conditions if the vast majority of other samples are attaining.

This is true especially for chronic criteria, where excursions that persist for less than a 4-day averaging period are not deemed to impair aquatic life.

This interpretation does not account for any increases in confidence with larger sample sizes. In a three-year period, 26,280 potential one-hour average concentrations and 274 potential four-day average concentrations can be measured with a perfect monitoring program. Two exceedances translates to a 0.00076% exceedance rate of all possible acute samples and 0.72% exceedance rate of all possible chronic samples. In other words, more than 99.2% of samples would have to be below the chronic criteria to be considered attaining the water quality standard. This level of certainty is not realistically attainable with the amount of data available to most assessment programs.

Actual assessment sample sizes vary due to monitoring program sampling frequency (e.g. weekly or monthly) and monitoring duration (e.g. a site is sampled for 6 months, 2 years or 10 years). Observed exceedance rates vary according to sample size. Two exceedances out of the minimum two-sample size for listing is a 100% exceedance rate. Two exceedances out of quarterly samples collected for 3 years is a 16.6% exceedance rate. Using a two-sample threshold as the critical exceedance rate results in an inconsistent and low level of confidence in both Category 2 and Category 5 listings – with the probability of listing dependent on the total number of samples collected.

The Raw Score Method

The raw score method uses a fixed percentage of allowable excursions out of the total number of samples collected as the critical exceedance rate. Rates commonly used by state assessment programs vary from 5%-10% for chronic toxic substances and 10% - 25% for conventional pollutants. With 10% the most commonly used (see Appendix 2, Table 7, below). If the proportion of samples collected is above the criterion magnitude is greater than the critical exceedance rate within in a 3-year period, the waterbody exceeds the standard.

The raw score of 10% essentially compares the 90th percentile of the observed sample concentration against the criterion magnitude. This exceedance rate matches the 7Q10 flow exceedance probability cited by EPA in the establishment of the criteria frequency component of the standards. This is approximately 10% in any given year[8].

The raw score is widely used by states to determine attainment of criteria for conventional pollutants like dissolved oxygen and pH. The rationale for the raw score method is that aquatic life can recover from brief, infrequent excursions above criteria magnitudes without detrimental effect– especially if the stressor is a naturally occurring compound or process[8]. The expectation is that organisms are adapted to periodic excursions of these parameters that occur even under natural conditions.

[8] EPA 1985, Technical Support Document for Water Quality based Toxics Control. EPA-440/4-85-032, United States Environmental Protection Agency. September 1985.

The raw score approach scales with the sample size. It does not scale for sample sizes less than 10, where one sample would represent >10%. It assumes that the samples evaluated represent the true distribution of pollutant concentrations in the waterbody. It assumes that the data is normally distributed and that the sample at hand represents the true population distribution of the water quality data in the waterbody.

The confidence in whether the highest 10% of samples represents the true 90th percentile of pollutant concentration in the waterbody depends on the total number of samples collected. With small sample sizes, uncertainty and the probability of making type-I (false-positive) errors is high and decreases with larger sample sizes. [9,10] As sample sizes increase, the distribution of the samples will better estimate the true population distribution of the waterbody.

As the size of a data set changes, the number of samples above the 90th percentile range change as well. The 90th percentile interval for a sample of 12 monthly samples is 0.9*12 = 10.8 or ~10 samples; no more than 2 samples can exceed the criterion. Increase the sample size modestly to 50, and the number of samples within the confidence interval is 45, and 5 samples are above the 90th% percentile. Although widely used, this method is not strongly recommended because it assumes the sample collected represents the true population variability of all possible samples in the waterbody, and the confidence level and error rates are strictly a function of the total number of samples collected.[9,10]

Table 1. Example 10% Raw Score Critical Values

Raw Score 10% critical exceedance frequency with 2 sample minimum

Sample Size	List if excursions are greater than:
1-10	1
11-20	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

Statistical Hypothesis Tests

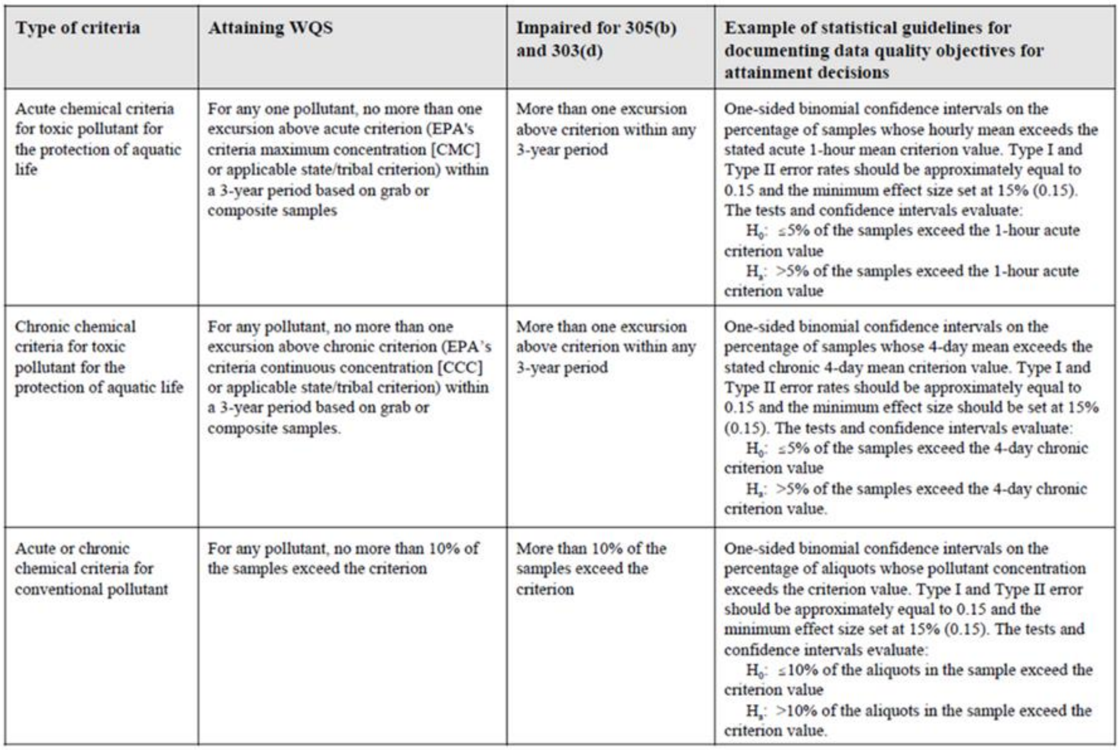
[9] Gibbons, 2003. A Statistical Approach for Performing Water Quality Impairment Assessments.

[10] Smith 2001. Statistical Assessment of Violations of Water Quality Standards under Section 303(d) of the Clean Water Act. Environmental Science and Technology, 2001 (34) 606-612.

Evaluating waterbodies with samples collected implies that the characteristics of the waterbody are accurately represented by the samples. A grab sampling inherently introduces bias, error, variability, and uncertainty about how well the samples represent the waterbody as a whole. Statistics test allow us to test the validity of the sample and provide a means to quantify whether a standard is being met.

Statistical analysis of water quality samples provides a quantifiable way to describe the confidence that a waterbody attains or exceeds a water quality criteria (which is comprised of magnitude, duration and frequency metrics) based on the samples collected. The EPA provides for a number of preferred statistical approaches to evaluate attainment of standards for different types of pollutants and parameters (
Figure 1). [12]

Figure 1. EPA example statistical guidelines for determining data quality objectives for attainment decisions.[11]



Type of criteria	Attaining WQS	Impaired for 305(b) and 303(d)	Example of statistical guidelines for documenting data quality objectives for attainment decisions
Acute chemical criteria for toxic pollutant for the protection of aquatic life	For any one pollutant, no more than one excursion above acute criterion (EPA's criteria maximum concentration [CMC] or applicable state/tribal criterion) within a 3-year period based on grab or composite samples	More than one excursion above criterion within any 3-year period	One-sided binomial confidence intervals on the percentage of samples whose hourly mean exceeds the stated acute 1-hour mean criterion value. Type I and Type II error rates should be approximately equal to 0.15 and the minimum effect size set at 15% (0.15). The tests and confidence intervals evaluate: H_0: ≤5% of the samples exceed the 1-hour acute criterion value H_a: >5% of the samples exceed the 1-hour acute criterion value
Chronic chemical criteria for toxic pollutant for the protection of aquatic life	For any pollutant, no more than one excursion above chronic criterion (EPA's criteria continuous concentration [CCC] or applicable state/tribal criterion) within a 3-year period based on grab or composite samples.	More than one excursion above criterion within any 3-year period	One-sided binomial confidence intervals on the percentage of samples whose 4-day mean exceeds the stated chronic 4-day mean criterion value. Type I and Type II error rates should be approximately equal to 0.15 and the minimum effect size should be set at 15% (0.15). The tests and confidence intervals evaluate: H_0: ≤5% of the samples exceed the 4-day chronic criterion value H_a: >5% of the samples exceed the 4-day chronic criterion value.
Acute or chronic chemical criteria for conventional pollutant	For any pollutant, no more than 10% of the samples exceed the criterion	More than 10% of the samples exceed the criterion	One-sided binomial confidence intervals on the percentage of aliquots whose pollutant concentration exceeds the criterion value. Type I and Type II error should be approximately equal to 0.15 and the minimum effect size set at 15% (0.15). The tests and confidence intervals evaluate: H_0: ≤10% of the aliquots in the sample exceed the criterion value H_a: >10% of the aliquots in the sample exceed the criterion value.

The EPA requirements for states wishing to use a statistical method to evaluate critical exceedances are that they must:

- provide a statistically verifiable level of assurance that the standard is attained.

[11] EPA, 2002. EPA, 2002. Consolidated Assessment and Listing Methodology (CALM). Toward a Compendium of Best Practices, First Edition. United States Environmental Protection Agency. July 2002 Section 4–6.

- explain selection of key sample statistics (arithmetic mean concentration, median concentration, or a percentile) to represent the critical exceedance rate.
- use an appropriate null and alternative hypothesis for listing and delisting decisions.
- quantify the assumed error rate / uncertainty expected in assessment data.
- define thresholds for type-I and type-II error rates and demonstrate they are effectively managed.
- provide a clear explanation of which statistical or analytical tools the state uses and under which circumstances.

For evaluating samples using a statistically-based methodology, definition of an alternate critical exceedance rate that is non-zero is required. EPA has endorsed, in numerous guidance documents, acceptable statistical alternatives to using the simple two-sample excursion as the critical exceedance rate to evaluate the frequency component of the water quality standards (Table 2). The selection of a critical exceedance rate other than two sample excursions is not a change to water quality standards, because it does not change the criteria being evaluated Instead, the critical exceedance rate is only used to quantify the strength or persuasiveness of the data used to interpret numeric water quality standards. It does not justify allowing the waterbody to exceed the standard some additional percentage of the time, as this would be an inappropriate interpretation of the frequency component[12].

The range of critical exceedance rates is mainly determined by EPA guidance. The standard critical exceedance rates are 5% for toxic substances and 10% for conventional pollutants. These rates are considered to provide a comparable level of protection to reflect the duration and frequency component of water quality criteria that are established in water quality standards. As such, DEQ anticipates it has limited ability to adopt different critical exceedance rates– unless they are recommended to be more stringent. A 5% rate for toxic substances and a 10% rate for conventional pollutants is almost universally applied by other states.

[12] EPA 2008. United States Environmental Protection Agency Determination Upon Review of Amended Florida Administrative Code Chapter 62-303. Identification of Impaired Surface Waters. Appendix A. Binomial Statistical Test. February 13, 2008.

Table 2. Critical Exceedance Rates for Toxic Substances Endorsed by EPA*

Critical Exceedance Rate	Source	Application
≤ 1 sample in 3 years	EPA, 1997**	Fully supports beneficial uses for acute criteria
0.09% (1 sample out of 1,095)	EPA, 2002	Hypergeometric distribution equivalent to a 1 in 3 year frequency of daily averages. For acute criteria.
0.36% (1 sample out of 274)	EPA, 2002	Hypergeometric distribution equivalent to a 1-in-3 year frequency of 4-day averages. For chronic criteria.
5% plus a 15% effect size	EPA, 2002	For toxics criteria, equivalent to a 1-in-3-year frequency.
10%	EPA, 2003	For chronic criteria, and acute criteria if justified, using a binomial or raw score test.
>10% raw score	EPA, 1997	For acute criteria not supporting beneficial uses. Sampling and measurement error accounted for.

*Adapted from Cal-EPA 2004. California Environmental Protection Agency, State Water Resources Control Board, Division of Water Quality, Functional Equivalent Document, Water Quality Control Policy for Developing California's Clean Water Act Section 303(d) List. September 2004.

**Information contained in the most recent EPA guidance for assessment, listing, and reporting requirements was intended to supersede previous guidance.

EPA, 1997 Guidelines for preparation of the comprehensive state water quality assessments (305(b) reports) and electronic updates: Supplement. EPA-841-B-97-002B
EPA, 2002. Consolidated assessment and listing methodology. Toward a compendium of best practices. First edition.
EPA, 2003 Guidance for 2004 assessment, listing and reporting requirements pursuant to sections 303(d) and 305(b) of the Clean Water Act.

Confidence Levels

The confidence value represents the desired certainty that small sample sizes are truly representative of the entire population. For example, when estimating the proportion of samples that would be exceeding in the waterbody from a set of water quality samples, one tests whether the confidence interval for the sample contains the critical exceedance rate. If the critical exceedance rate is less than the confidence interval for the sample, then the waterbody can be considered impaired. The confidence level directly determines the probability of making a type-I error. For a 90% confidence interval, there is a 10% chance of making a type-I error for any given application of the test. As confidence levels go down, the probability of making a type-I error goes up.

There is no objective method for selecting an ideal confidence interval. Selection of a confidence interval is done *a priori* as a condition of the statistical test. Statistically valid confidence intervals for hypothesis testing range from 80% -99%. The standard confidence level for scientific research is 95%. For regulatory purposes, 80%- 90% may be sufficient. A range of 80%-95% percent is used for assessment in other states. Only California and Texas apply a confidence level of 80% to statistical assessment methods. California applies the 80% confidence level, but this is offset by adoption of a more stringent critical exceedance rate of 3%. Texas varies the confidence level according to sample size, which can range from 80% – 40%. Nine other states that apply a statistical method to assessment use a confidence level of 90%. EPA recommends a 90% confidence level in guidance.

For water quality assessment, the selection of a confidence interval will directly affect the number of listings that will result. Choosing a confidence interval range is an inherently subjective process. DEQ seeks to avoid both unnecessary economic and opportunity costs incurred to the regulated community and DEQ's TMDL program by overestimating the number of impaired waters through type-I errors. DEQ also seeks to avoid unnecessary costs to the environment and beneficial use of waters incurred by failing to identify impaired waters through type-II errors. DEQ has determined that a 90% confidence level is expected to balance program needs for accuracy while remaining consistent with EPA guidance and best practices.

Effect Size

The critical exceedance rate of samples indicating impairment (Category 5) can be different from the number required to find attainment (Category 2). EPA guidance recommends a statistical guideline of an acceptable exceedance frequency of 10% (on average) and an unacceptable exceedance frequency of 25% in any given sample. To address the different tolerance for risk of listing a waterbody that ***is not*** actually impaired versus ***not*** listing a waterbody that ***is*** actually impaired. This is referred to as "balancing" the error rates. The difference between these critical exceedance rates is called the effect size.

Effect size is the maximum proportion of exceeding samples that would be tolerated before listing. A "regulatory" critical exceedance rate of 10% and an "unacceptable" critical exceedance rate of 25% reflects an effect size of 15%. Following EPA's guidance, waters with less than a 10% proportion of excursions would be considered attaining the standard. Waters with an exceedance frequency above 25% would always be determined to be impaired, and placed on the section 303(d) list.

Waters with a proportion that fall between these two values would sometimes be listed and sometimes not. The probability of making a false Category 5 decision is determined by the chosen confidence level of the statistical test (the type-I error rate). The probability of making a false Category 2 decision is determined by type-II error rate. While the lower or “regulatory” critical exceedance rate (i.e. 10%) determines whether a waterbody is placed into Category 2 or Category 5, the higher “unacceptable” exceedance rate can be used to calculate the probability of making a type-II error.

Oregon DEQ’s assessment of numeric criteria

Many states face challenges assessing monitoring data to determine whether a relatively small number of samples indicates a waterbody attains or exceeds the duration and frequency components of water quality criteria[13]. States using the >1-sample-in-3-years as a critical exceedance rate also face challenges assessing large data sets. The question is whether one or two individual sample excursions are representative when the overwhelming number of samples in a large data set show attainment. Oregon DEQ’s Ambient Monitoring Network collects samples from sites six times a year (once every 2 months). Financial constraints limit the likelihood of aligning the frequency of sampling to directly measure the 96-hour and 1-hour exposure duration.

Oregon’s 2010 Assessment Methodology followed the raw score method and allowed for listing on two sample excursions as long as these represented 5% or more of the proportion of samples. However, DEQ did not assess toxic substances in the 2010 assessment cycle, so this methodology was never applied to listings.

Oregon’s 2012 303(d) Integrated Report reverted to the >1-sample-in-3-year excursion method with a critical exceedance rate of >1 excursion. This was applied to assess the most stringent of the aquatic life toxics or human health criteria with a minimal sample size of only two. Listing a waterbody in Category 5 required just two sample excursions over the magnitude. The minimal sample size required to assign a waterbody to Category 2 was five samples. DEQ acknowledged in their *2012 Response to Comments* that future assessment methodology could consider protocols to evaluate large data sets and apply the frequency and duration elements of the aquatic life and human health criteria. Because the most stringent criteria are usually toxic substances criteria, the human health criteria were almost entirely unassessed.

For conventional pollutants, Oregon continued to apply the raw score method with a 10% critical exceedance rate in 2012. Most conventional pollutants used a critical exceedance rate of >1-sample or 10%, whichever was greater. A minimum of five samples was required to place a waterbody in Category 5. A minimum of 10 samples was required to place waterbodies in Category 2. Waterbodies with only 6–9 samples were placed in Category 3. These parameters include pH, dissolved oxygen, bacteria, and total dissolved gas.

[13] Kansas Department of Public Health & Environment, 2011. Duration and Frequency for Assessing Numeric Criteria. WQ Standards White Paper.

Significant public comment was received on the 2010 and 2012 draft 303(d) lists citing the listing methodology for toxic substances. Twenty-six percent of the 111 listings for toxic substances on Oregon’s 2012 303(d) list were for only 2 sample excursions, all for chronic criteria. The 3-year frequency of excursions was not evaluated. The median sample size for a listing was 24 samples.

Historically, in the absence of robust data sets and in an effort to be protective of beneficial uses, DEQ has erred on the conservative side, using a low threshold for placing waterbodies in Category 5 on the state’s 303(d) list. The time to complete an individual TMDL ranges from 24 months to 5 years. The state must have a means to target limited resources to identify and address the most critical impairments from data that represent the true and current condition of waterbodies.

Table 3. Summary of critical exceedance rates for numeric criteria from the 2012 Assessment Methodology

	Chronic			Acute	
	Attaining	Exceeding	Confidence Level	Attaining	Exceeding
Aquatic Life Toxics	≤1-sample	>1-sample	NA	≤1-sample	>1-sample
Conventional Pollutants	Raw score ≤10%	Raw score >10%	NA	NA	NA
Human Health Criteria	≤1-sample	>1-sample	NA	NA	NA

Alternative Policy Options for Assessment Methodology of Aquatic Life Toxics Criteria

With this effort, DEQ is closely examining how it evaluates data for making assessment conclusions. The goals that DEQ wants to achieve by reviewing this assessment methodology are:

1. Quantify and reduce the probability of making listing decision errors
2. Develop a methodology that accurately assesses Oregon waterbodies with a higher confidence in assessment conclusions

One challenge of an assessment methodology is to determine what constitutes an “impaired” water for purposes of listing. This issue is not unique to Oregon, and multiple states have taken different approaches to address this issue. The tools employed by other states to address large data sets are the use of a proportional critical exceedance rate (the raw score approach) or a statistical approach (i.e. binomial test) to account for variability in the representativeness and distribution of samples and sampling and analytical errors. See a review of other states' statistical listing methods in Appendix 2. There are several alternative policy options DEQ could adopt to assess surface waters for impairment of aquatic life due to toxic substances (Table 4).

Table 4. Alternative Listing Policy Options for Oregon
1. Status Quo: Utilize a two-sample critical exceedance rate with no minimal sample size for toxics and a 10% raw score for conventional pollutants.
2. Retain the status-quo exceedance rates but increase the minimum required sample size.
3. Use an alternate critical exceedance rate based on the raw score approach at 5% for toxic substances and 10% for conventional pollutants.
4. Use a statistical hypothesis test based on the binomial method for toxic substances and conventional pollutants.
5. Apply an alternate critical exceedance rate derived from a raw score or binomial approach to large sample sizes only (i.e. >18).
6. Apply a critical exceedance of two samples to acute toxics criteria and a critical exceedance based on a raw score or binomial approach to chronic criteria.

Policy Analysis

Option 1: Status quo. Maintain a two-sample critical exceedance rate for toxic substances and a 10% raw score for conventional pollutants.

In its 2012 Assessment Listing Methodology, DEQ used a non-statistical critical exceedance of two samples to list waterbodies as impaired. If any two samples were measured greater than the magnitude of the criterion the waterbody was considered to exceed the standard for that period. The method applied did not take into account the 3-year frequency period.

This literally interpreted the >1-in-3-year frequency and applied it to individual sample excursions. While this interpretation of water quality criteria has a very low type-II (false negative) error probability, it comes at the expense of an extremely high type-I (false positive) error probability[14]. The probability of making a listing error also increases with

[14] Smith 2001. Statistical Assessment of Violations of Water Quality Standards under Section 303(d) of the Clean Water Act. Environmental Science and Technology, 2001 (34) 606-612.

sample size. This was to ensure that the type-II error probability (false-negative) is almost zero. Because of sampling uncertainty and error, a more robust critical exceedance rate than two samples is justified.

The method of listing for any two excursions is overly conservative for large data sets or over longer periods (i.e. the 10 years that will be used for the 2018 cycle versus 2 years for a typical Integrated Report). Errors made in the assessment process may be costly for both DEQ and permittees. Because of a “false positive” listing, additional regulatory restrictions are placed on NPDES permit holders who discharge to listed waterbodies. This increases the regulatory burden and cost for permit compliance for permitees and increases demands on DEQ’s resources for permit implementation and TMDL development without resulting in clear environmental benefits. Both type-I and type-II errors may be avoided by assessing the water quality situation more completely. Strictly speaking, if more monitoring data were available to better assess water quality conditions, then errors could be minimized. The cost of minimizing these errors is the cost of performing the monitoring.

Listings based on two exceedances, without consideration for the total number of samples does not fully utilize data generated from a robust ambient monitoring program and serves as a disincentive for entities to submit their large ambient datasets to DEQ for use in the assessment.

Using the two sample critical exceedance rate also assumes that instantaneous grab samples are representative of the true average concentration of 1-hour or 96-hour duration. EPA reserves for states to determine how far to extrapolate instantaneous single grab samples to conditions in time.[15]

Advantages:

1) Simple to implement and interpret.
2) It is likely that an individual grab sample adequately represents a 1-hour average concentration for evaluating acute criteria.
3) Very low type-II (false negative) error probability of failing to list a waterbody if the data represents a true impairment.

Risks:

1) It is not likely that a monthly or quarterly instantaneous grab sample adequately represents a 4-day average concentration; adds uncertainty to assessment of chronic criteria.
2) No way to specify a confidence level for either small or large data sets.

[15] EPA, 2005.

3) Poorly suited to assessing large data sets.
4) Assumes that as few as two samples adequately represents the variability in the waterbody.
5) Very high and completely uncontrolled type-I (false positive) error probability.
6) Regulatory and economic burden of erroneous Category 5 listings.
7) Disincentive for independent monitoring and data sharing by third parties.

Option 2: Retain the two-sample critical exceedance rate and increase the minimal sample size for listing

Increasing the minimal sample size required to list a waterbody in Category 5 from two samples to 10 increases the potential confidence of the assessment. However, maintaining the two-sample critical exceedance rate does not address potential variability and error in the representativeness of samples. It shares the same advantages and risks of option 1.

Advantages:

1) Slight increase in the confidence of listing and attainment decisions.
2) Slight reduction in type-II error probability for impairment.

Risks:

1) Same as status quo.

Option 3: Apply a 5% raw score critical exceedance rate for toxic substances and a 10% raw score for conventional pollutants.

In Oregon’s 2010 Assessment Methodology, DEQ called for a raw score method with a 5% critical exceedance rate for assessing acute and chronic toxic pollutant criteria. If ≥5% of the samples were higher than the magnitude, the waterbody was determined to be impaired.

This is equivalent to comparing the 95th percentile of sample concentration to the magnitude of the water quality criterion. With this method, the number of allowed excursions changes with sample size. However, the method assume the proportion of excursions observed in the sample exactly matches the proportion of excursions in the waterbody.

While it better accounts for the 90% confidence interval by considering the number of samples collected at a site, the raw score method also tends to have a high false-positive rate[16,17].

[16] Smith et al., 2001. Statistical Assessment of Violations of Water Quality Standards under Section 303(d) of the Clean Water Act. Environmental Science and Technology 35 (2001) 606-612.

Advantages:

1) Simple to implement and interpret.
2) Confidence scales with the number of samples representing conditions in the waterbody.
3) Provides some control over Type-I (false-positive) error rates.
4) Makes some allowance for natural variability and sampling error.
5) Recognizes short-term excursions are not likely to harm aquatic life.
6) Less of a disincentive for collection of larger, more representative sample sets.

Risks:

1) Not as robust as hypothesis-based methods.
2) Has a relatively high type-I (false-positive) error probability relative to sample size.
3) Does not directly address Type-II (false-negative) error probability.
4) Low confidence in sample sizes <10.

Option 4: Adopt a hypothesis test based on the binomial method

The exact binomial test is currently the most widely used hypothesis test and has been adopted in at least nine states (see Appendix 2). The binomial distribution methodology concept is to statistically determine, with a desired level of certainty, whether the observed number of excursions that exceed a pollutant concentration limit in a set of random samples of the waterbody, would be likely to occur if the proportion of samples in the waterbody as a whole was greater than the critical exceedance rate. The binomial test as applied to water quality assessment is used to determine what proportion of the sample could exceed the magnitude of the criterion before the waterbody as a whole is considered impaired. The binomial test has been described as a modest improvement in controlling both type-I and type-II error over the EPA raw score method.[18]

The binomial test is based on how well samples match an expected distribution of yes/no or pass/fail outcomes. Each water sample is evaluated for whether it is above or below the criterion concentration threshold. The method is based on defining the number of samples that can exceed the threshold *in the population* if they were to collect all possible samples of the waterbody, and still be considered attaining water quality standards. The more samples that are collected, the greater the confidence in the conclusion about the waterbody.

[17] Gibbons, 2003. A Statistical Approach for Performing Water Quality Impairment Assessments. Journal of the American Water Resources Association 39(4): 841-849

[18] Smith et al, 2001. Statistical Assessment of Violations of Water Quality Standards Under Section 303(d) of the Clean Water Act. Environmental Science and Technology 35 (2001): 606-612.

The test uses the binomial distribution, a discrete distribution of the proportion of the number of "attaining" and "exceeding" measurements in a sample, to determine whether the *subset of pollutant samples actually collected* would indicate the true proportion of *all possible samples* in the waterbody also exceed the criterion threshold.

EPA guidance for applying the binomial test recommends either a 5% or 10% critical exceedance rate, and a 90% confidence level. The recommended null hypothesis for listing would be that the actual exceedance in the population is ≤5%. The alternative hypothesis is that the actual exceedance rate is >5%. Some states (e.g. California, Washington prior to 2014) also incorporate the recommended effect size (See Appendix 1) of 15% into the binomial test by setting the exceedance rate of the alternative hypothesis 15% higher than rate of the null hypothesis.

Since there is greater risk in assuming an impaired waterbody is not impaired, the null hypothesis for delisting is that we assume the waterbody is not attaining, and more than >10% of samples are above the magnitude of the criterion. With smaller sample sizes, it is harder to reject the null hypothesis and we are more likely to keep a waterbody on the list if there are any exceedances. However, there is a minimal sample size necessary to achieve a given confidence level for two excursions.

Figure 2. The number of exceedances that determine impairment and sample size ranges using the binomial method for a 5% critical exceedance rate for 90% and 80% confidence.

Critical Values for Listing Chronic Toxic Substances		
Sample Size		
90% Confidence	80% Confidence	List if excursions ≥ :
5-18	5-18	2
19-22	19-30	3
23-35	31-46	4
36-49	47-62	5
50-63	63-78	6
64-78	79-95	7
79-92	96-112	8
93-109	113-129	9
110-125	130-146	10
126-141	147-164	11
142-158	165-181	12
159-171	182-200	13
179-191	201-217	14
192-200	218-234	15

Benefits:

1) More complicated to calculate, but thresholds for the desired critical exceedance and confidence interval can be pre-calculated and presented as lookup tables.
2) Treats water quality samples as samples of the waterbody (population)[19]
3) Estimates the "true" exceedance given the number of exceedances observed and the number of water samples actually collected.
4) Reduces type-I error probability over raw score and 1-in-3-year excursion methods.

[19] Smith et al, 2001. Statistical Assessment of Violations of Water Quality Standards Under Section 303(d) of the Clean Water Act. Environmental Science and Technology 35 (2001): 606-612.

EPA 2002, Consolidated assessment and listing methodologies (CALM). Towards a compendium of best practices. First Edition. July 2002.

EPA, 2005. Florida Impaired Waters Rule Appendix A: Detailed Review of Binomial Statistical Test

5) Ability to directly control the type-I error probability by selecting the desired statistical confidence level
6) The uncertainty in type-II error is directly quantified and can be adjusted through sample size, and selection of effect size and the confidence level.
7) Provides strong incentive for longer-term monitoring

Risks:

1) The type-II error probability is higher than for the >1-in-3-years critical exceedance rate
2) Attaining at least 90% confidence in a 10% critical exceedance rate requires a minimal sample size of 10 (Figure 2).

Option 5: Adopt an alternative critical exceedance rate based on the binomial method, but limit application to large sample sizes.

By maintaining the status quo of using a >1-in-3-year critical exceedance rate for small sample sizes, this approach reduces the type-II error for limited data sets, while limiting the type-I error for larger data sets. Smaller datasets are more prone to Type-II error under the binomial approach, while larger data sets are more prone to type-I error under the >1-in-3-year approach.

Advantages:

1) Generally same as Option 4.
2) Additional reduction in the type-II error rate for small sample sizes.
3) Provides strong incentive for additional monitoring by 3rd parties.

Risks:

1) Generally same as option 1 for small data sets.
2) Maintains same type-I error rate for small sample sizes as the >1-in-3-year approach.

Option 6: Apply a critical exceedance of ≥2 samples to acute criteria and a critical exceedance based on a binomial approach to chronic criteria

This option maintains the status quo for listing a waterbody on >1-in-3-year sample excursion for acute criteria, but applies a binomial test at an alternate critical exceedance rate to chronic criteria.

Advantages:

1) Generally the same as option 4.

2) Better aligns with the ecological risk of committing a type-II error for acutely toxic levels of pollutants.
3) Matches the average community recovery interval found in studies of response to significant disturbance
4) Greater likelihood that an instantaneous grab sample represents a 1-hour average versus a 4-day average concentration.

Risks:

1) Generally the same as Option 4.

Public Input for Methodology Updates

Under Oregon statute, (ORS 468.B.039) DEQ is required to release its assessment methodology for public comment prior to drafting the 303(d) / 305(b) report. Since this is a significant revision of the existing listing methodology, it was important that DEQ seek additional public input before finalizing the assessment approach.

DEQ conducted an external scientific peer review in January of 2018. The results of the peer review are presented in Appendix 4. regard to the statistical methodology proposed and the selection of parameters for the statistical tests. This includes choosing a suitable critical exceedance rate, formulation of appropriate null hypotheses, effect sizes, and confidence intervals for waterbody assessment.

Conclusion

Recommendation

It is DEQ's recommendation to adopt a combination of Option 5 and Option 6 as the method for listing waterbodies for impairment of water quality standards for toxic substances and conventional pollutants for the protection of aquatic life. We further recommend applying the geometric mean of samples to the human health criteria for the entire assessment window for lifetime exposure duration, or for geometric means that match the duration indicated in the water quality criteria (e.g. 30-day average concentration).

By using a binomial statistical approach and increasing the sample size, there is a greater likelihood of making an accurate impairment conclusion when sample sizes are large. Retaining the >1-in-3-year critical exceedance rate for small sample sizes, and acute criteria, addresses greater risk to aquatic life of type-II errors with these types of datasets. This simultaneously addresses the very high risk of type-I errors for larger datasets using the >1-in-3-year as a critical exceedance rate, and the slightly higher type-II error rate for the binomial method.

DEQ convened an independent peer review panel to review the selection of the null and alternate hypotheses, critical exceedance rates, and effect size for application of the binomial test in Oregon. The results of this peer review are presented in Appendix 4.

Proposed Listing Methodology

Acute Toxic Substances Criteria:

- Continue to apply >1-sample- in-3-year critical exceedance frequency
- Minimum sample size to list is 5 samples

Chronic Toxic Substances Criteria and Conventional Pollutants:

- Continue to apply the >1-in-3-year critical exceedance rate to all data sets with <18 samples.
- Apply a binomial test with a 5% or 10% critical exceedance rate at 90% confidence to determine impairment for the purpose of Category 5 determinations for sample sizes ≥18.
- Null hypothesis (H_0): that ≤5% or ≤10% of samples in the population exceed the criterion (attaining)
- The Alternative hypothesis (H_A) the actual exceedance is >5% or >10%.
- The effect size is 15%.
- Consult an independent technical review panel to confirm or adjust the proposed confidence intervals and null hypotheses.

Human Health Criteria

- Apply the geometric mean for the appropriate duration of all applicable samples within the assessment window to the criterion magnitude.
- A geometric mean greater than the criterion magnitude indicates impairment.

Table 5. Proposed Listing Methods for Numeric Criteria

	Chronic			Acute	
	Attaining	Exceeding	Min. confidence interval	Attaining	Exceeding
Aquatic Life Toxics	Binomial H_O: ≤ 5% of samples exceed the 4-day chronic criterion value	Binomial H_A: >5% of samples exceed the 4-day chronic criterion value	90%	≤1-sample-in-3-years	>1-sample-in-3-years
Conventional Pollutants	Binomial H_O: ≤ 10% of samples exceed the criterion value	Binomial H_A: >10% of samples exceed the criterion value	90%	NA	NA
Human Health Criteria	Geometric mean sample concentration ≤ criterion	Geometric mean sample concentration > criterion	NA	NA	NA

Figure 3. Minimal number of sample excursions to list as impaired following the proposed binomial procedure for toxic substances [20]

Null Hypothesis: Actual exceedance proportion is ≤5%
Alternate hypothesis: Actual exceedance proportion is >5%
Minimum confidence level is 90%

Sample Size	List if excursions ≥
2-18	2*
19-22	3
23-35	4
36-49	5
50-63	6
64-78	7
79-92	8
93-109	9
110-125	10
126-141	11
142-158	12
159-171	13
179-191	14
192-200	15

* sample sizes <18 use >1-samples-in-3-year critical exceedance rate

[20] Adapted from CA-SWRCB, 2004. California State Water Resources Control Board Water Quality Control Policy for Developing California's Clean Water Act Section 303(D) List. September 30, 2004.

Figure 4 Cumulative error probabilities for the toxic substances listing procedure. 5% critical exceedance rate, 90% confidence level.

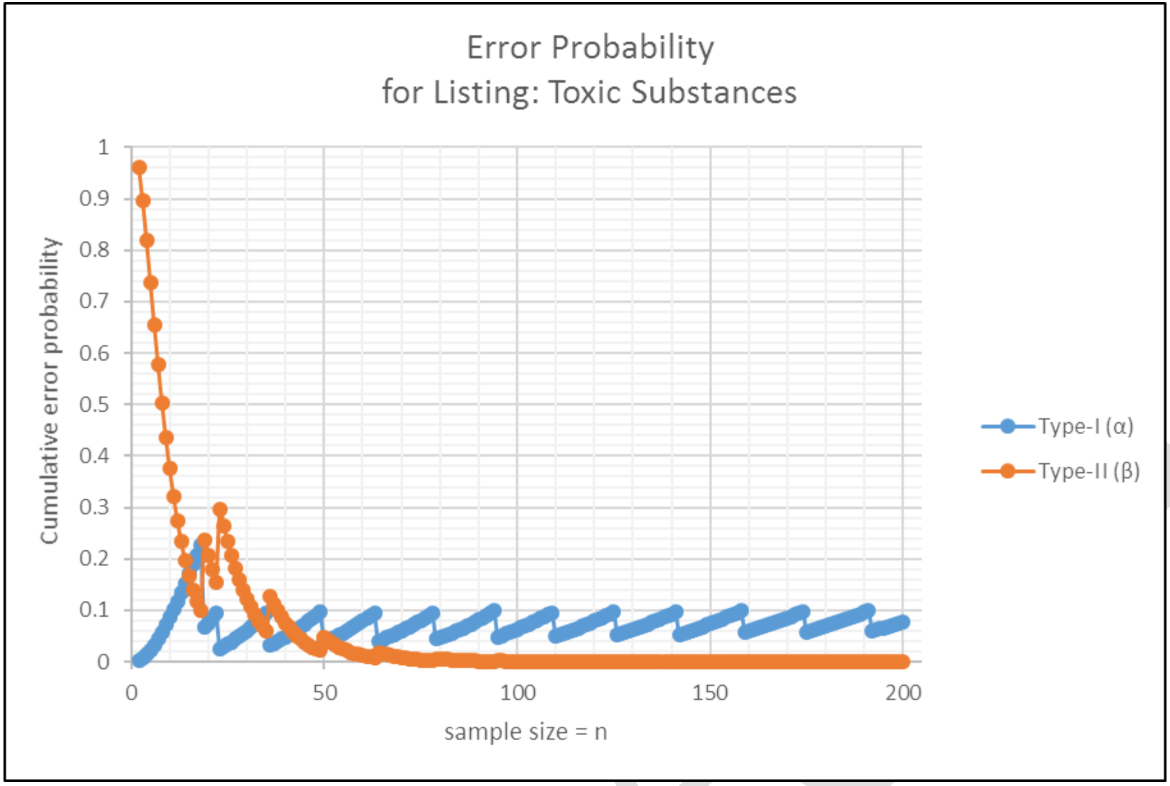


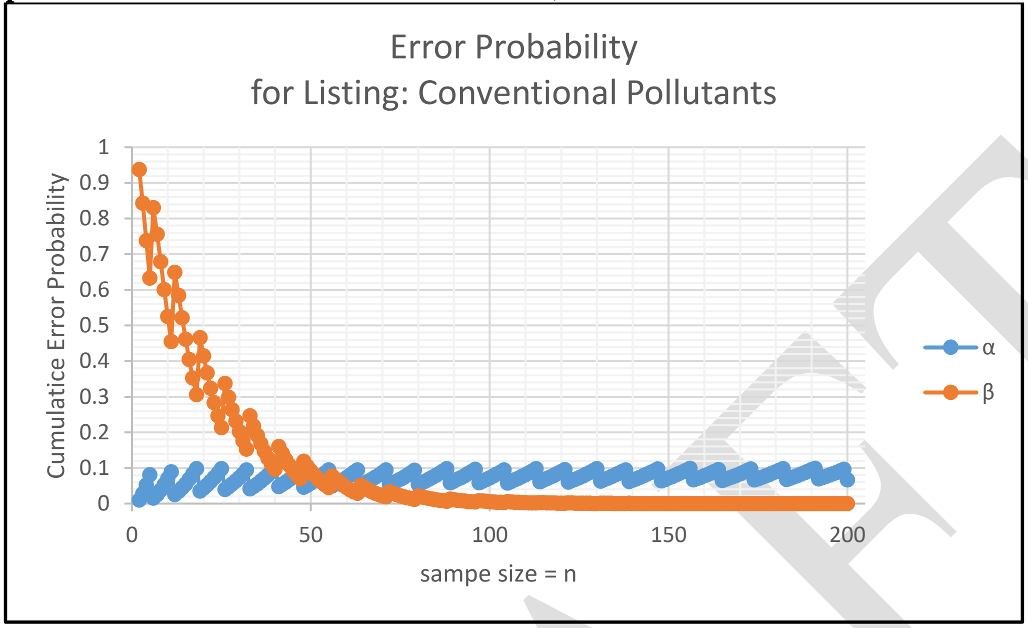
Figure 5. Example of minimal number of sample excursions to list as impaired following the proposed binomial method for conventional pollutants [21]

Null Hypothesis: Actual exceedance proportion is ≤10%
Alternate hypothesis: Actual exceedance proportion is >10%
Minimum confidence level is 90%

Sample Size	List if excursions ≥
10-18	4
19-25	5
26-32	6
33-40	7
41-47	8
48-55	9
56-63	10
64-71	11
72-79	12
80-88	13
89-96	14
97-104	15
105-113	16
114-121	17

[21] Adapted from CA-SWRCB, 2004. California State Water Resources Control Board Water Quality Control Policy For Developing California's Clean Water Act Section 303(D) List. September 30,2004.

Figure 6 Cumulative error probabilities for the conventional pollutant listing procedure. 10% critical exceedance rate, 90% confidence level.



Alternate Recommendation:

Apply the raw score method to conventional pollutants at a critical exceedance rate of 10%, chronic toxic substances at 5% as was proposed for the 2010 methodology, acute toxic substances at >1-sample-in-3-years, and geometric mean for the human health criteria.

Potential Impact

In the past, Oregon has assessed against the most stringent criterion only. The recent 303(d) list has a majority of Category 5 listings for excursions of the chronic, and not acute, criteria. Incidental excursions (<5%–10%) above chronic criteria are not expected to be detrimental with the same severity as excursions above acute criteria. This justification was accepted by EPA Region 4 for Florida's adoption of the binomial method[22].

As shown in Figure 4 and Figure 5, the selection of the 90% confidence interval means that the probability of falsely listing a waterbody (α, or type-I error) is kept to 10% or less. The probability of failing to list an impaired waterbody (β, or type-II error) is significantly reduced as sample sizes increase. Type-II error probabilities were calculated based on the 15% effect size. The procedure for calculating critical rates and error probabilities is reported in Appendix 3. The inflated type-II error probability for small

[22] EPA 2008. United States Environmental Protection Agency Determination Upon Review of Amended Florida Administrative Code Chapter 62-303 Identification of Impaired Surface Waters. Appendix A. Binomial Statistical Test. February 13, 2008.

sample sizes, related to a failure to list a waterbody that may actually be impaired, is addressed by ongoing monitoring. Expanding the time period for data in the assessment, to evaluate data from more than the most recent data collected since the last assessment, would also result in larger sample sizes and reduced type-II error.

DEQ surveyed Category 5 listings in the Willamette Basin for toxic substances for the protection of aquatic life in the draft 2012 303(d) list. Since DEQ assessed numeric criteria against the most stringent criteria, 100% of listings were for exceedances of the chronic criteria. If DEQ had adopted Option 6, the binomial approach with a >1-sample-in-3-year critical exceedance rate for acute criteria the listings would be reflected in the following way:

1) 45% of Willamette Valley Listings were made on <18 samples and would not see a de-facto change in listing based on two exceedances for smaller data sets.
2) 26% of Category 5 determinations were made on the basis of only 2 sample excursions.
3) For 7% of the Willamette Category 5 listings, the number of exceeding samples was < 1% of the total sample size.
4) 4% of the listings were based on a total sample size of only 2 samples.
5) 17% of listings were for sample sets of 4 or less, equivalent to only one year of quarterly monitoring.
6) 17% of listings in the Willamette (15 of 111) would be reconsidered as attaining if applying a binomial method at 5% critical exceedance with a 90% confidence rate as presented in Figure 3.

Case Study #1:

Several examples of listings with large data sets were identified in comments received during the 2012 303(d) listing process. One example of comments received were from Clean Water Services regarding total recoverable copper listings at twelve locations on the Tualatin River.

Supporting data identifies 7 monitoring locations in the lower Tualatin River and 5 monitoring locations in the upper Tualatin River. In total, there are more than 600 valid data points in the lower Tualatin River and nearly 250 valid data points in the upper Tualatin River. Table 6 shows the monitoring locations in the upper and lower Tualatin River, the number of valid samples, the total number of exceedances, and the total percent of exceedances for copper in the Tualatin River.

The impaired waters listing determination resulted from applying the >1-sample-excursion as the critical exceedance rate for the toxics criteria. More than 2 sample excursions were observed at each monitoring location, resulting in the listing.

However, if the 5% raw score exceedance rate had been applied, only 3 sites in the upper watershed would be considered impaired. Although there were two or more exceedances of the criteria at all of the sampling sites, the frequency of exceedance for most sites was below five percent.

Applying the binomial test (Table 6) illustrates that even with 95% confidence, there are no sites in either the lower or upper Tualatin River where we expect the waterbody would exceed a critical exceedance rate of 10%.

Listing the Tualatin River as impaired for copper based on two or more exceedances does not constitute a reasonable approach when considering the information provided by the dataset. Although occasional exceedances of the copper criteria occur, it does not guarantee that they are representative of a 4-day average concentration for typical conditions in the waterbody. In addition, the assessment does not indicate whether or not these exceedances occurred on the same sampling date at multiple sites, ultimately reducing the total number of waterbody exceedances if they were spatially and temporally concurrent within the assessment unit.

Case Study #2

Another example that was identified in the comments on the 2012 303(d) list was a Category 5 listing for chromium in Gales Creek (Figure 7). Clean Water Services indicated that the listing for chromium was based on two exceedances out of a total of 54 samples. Only 15 of the samples were above the method detection limit. Although DEQ does not evaluate flow data when evaluating chemical data for the 303(d) assessment, the periodic pulses in chromium concentration are likely associated with short term events like storm flows. The potential that the duration of the excursions is greater than 4-days is not evaluated in the data, but an error rate of 5% would reasonably account for this possibility. This listing would not be found impaired under the proposed binomial assessment method of 5% critical exceedance at 90% confidence (6 samples, Table 7).

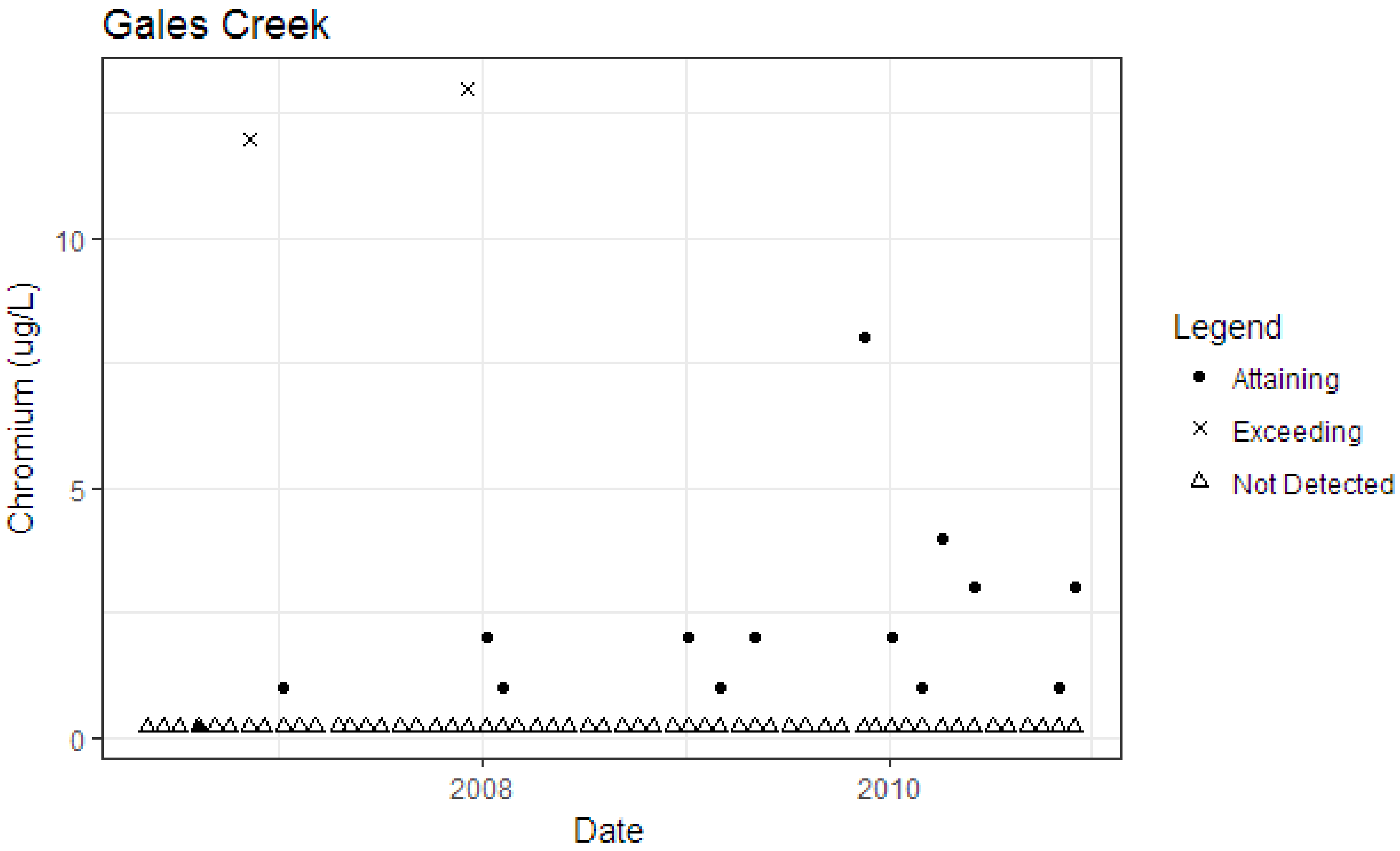
Table 6. Total Recoverable Copper in the Tualatin River (2002-2010)

Watershed	LOCID	Location	Valid Samples	Total Number Exceeded	Percent Exceeded (Raw Score)	Binomial number to list under 5% critical exceedance (90% Confidence	Binomial number to list under 5% critical exceedance (80% Confidence)	Remain Listed with Binomial Distribution	Remain Listed with Raw Score (5%)
Lower Tualatin River	3701002	TR @ Weiss Bridge	97	4	4.1%	9	8	NO	NO
	3701054	TR @ Stafford Road	98	3	3.1%	9	8	NO	NO
	3701087	TR @ Boones Ferry	98	4	4.1%	9	8	NO	NO
	3701165	TR @ Elsner	98	4	4.1%	9	8	NO	NO
	3701271	TR @ Scholls	97	4	4.1%	9	8	NO	NO
	3701333	TR @ Farmington	90	2	2.2%	8	7	NO	NO
	3701391	TR @ Rood Road	88	2	2.3%	8	7	NO	NO
Upper Tualatin	3701450	TR @ Hwy 219	97	2	2.1%	9	8	NO	NO
	3701528	TR @ Golf Course Road	96	6	6.3%	9	8	NO	YES
	3701569	TR @ Fernhill	36	3	8.3%	5	4	NO	YES
	3701612	TR @ Springhill	97	7	7.2%	9	8	NO	YES
	3701715	TR @ Cherry Grove	92	4	4.3%	9	7	NO	NO

Table 7. Chromium in Gales Creek Aquatic Life Criteria = 11 µg/L (Chromium VI)

Watershed	Location	Site Description	Valid Samples	Total Number Exceeded	Percent Exceeded	Number to list under binomial (5% exceed, 90% CI)	Number to list under 5% raw score (80th %-ile)	Remain Listed for Binomial	Remain Listed for raw score 5%
Tualatin	USGS-453229123101101	Gales Creek	54	2	3.7%	6	5	NO	NO

Figure 7. Chromium in Gales Creek. The measured samples (circles) show detectable concentrations for dissolved Chromium (III and VI). Symbols shaded red exceed the Chromium VI chronic criterion for aquatic life of 11 ug/L. Samples below the detection limit are shown as black crosses. n=54 samples.



References:

1. Barnett V, O'Hagan A. 1997. Setting environmental standards: The statistical approach to handling uncertainty and variation. First Edition. Chapman & Hall. New York. ISBN:9780412826207
2. EPA 1985, Technical Support Document for Water Quality based Toxics Control. EPA-440/4-85-032, United States Environmental Protection Agency. September 1985.
3. EPA 1994, Water Quality Standards Handbook: Second Edition. Appendix H. Derivation of the 1985 Aquatic Life Criteria. Environmental Protection Agency, Office of Water. EPA 823-B-94-005a. August 1994.
4. EPA, 1997 Guidelines for preparation of the comprehensive state water quality assessments (305(b) reports) and electronic updates: Supplement. EPA-841-B-97-002B https://www.epa.gov/tmdl/integrated-reporting-guidance
5. EPA, 2002. Consolidated Assessment and Listing Methodology (CALM). Toward a Compendium of Best Practices, First Edition. United States Environmental Protection Agency. July 2002. https://www.epa.gov/waterdata/consolidated-assessment-and-listing-methodology-calm
6. EPA, 2003. Guidance for 2004 assessment, listing and reporting requirements pursuant to sections 303(d) and 305(b) of the Clean Water Act. https://www.epa.gov/tmdl/integrated-reporting-guidance
7. EPA 2005a, Memorandum: Guidance for 2006 Assessment, Listing and Reporting Requirements Pursuant to Sections 303(d), 305(b) and 314 of the Clean Water Act. July 29, 2005. http://www.epa.gov/owow/tmdl/2006IRG
8. EPA, 2005b. Florida Impaired Waters Rule Appendix A: Detailed Review of Binomial Statistical Test.
9. EPA, 2008. United States Environmental Protection Agency Determination Upon Review of Amended Florida Administrative Code Chapter 62-303 Identification of Impaired Surface Waters. Appendix A. Binomial Statistical Test. February 13, 2008. https://archive.epa.gov/pesticides/region4/water/wqs/web/pdf/epa_iwr_decdoc_2-19-08.pdf
10. Gergs, Andre, Silke Classen , Tido Strauss , Richard Ottermanns , Theo C. M. Brock , Hans Toni Ratte , Udo Hommen , and Thomas G. Preuss, 2016. Ecological Recovery Potential of Freshwater Organisms: Consequences for Environmental Risk Assessment of Chemicals. Reviews in Environmental Contamination and Toxicology. 2016 (94) 236: 259-294. DOI:10.1007/978-3-319-20013-2_5.
11. Gibbons, Richard. 2003. A Statistical Approach for Performing Water Quality Impairment Assessments. Journal of the American Water Resources Association 39(4): 841-849 DOI: 10.1111/j.1752-1688.2003.tb04409.x
12. Harcum, Jon B. and Steven A. Dressing. 2015. Technical Memorandum #3. Minimal Detectabel Change and Power Analysis. October 2015. https\\www.epa.gov/xxx/tech_memos.htm
13. Kansas Department of Public Health & Environment, 2011. WQ Standards White Paper: Duration and Frequency for Assessing Numeric Criteria. www.kdheks.gov/water/download/tech/Duration_Frequency_final_Jan27.pdf

14. Keller, Arturo A. and Lindsey Cavallaro. 2008. Assessing the US Clean Water Act 303(d) listing process for determining impairment of a waterbody Journal of Environmental Management 86 (2008) 699–711.
15. Lin, P., M. Duane, and X.F. Niu. 2000. Nonparametric procedure for listing and delisting impaired waters based on criterion exceedances. Task l, Contract Number LAB015 Tallahassee, FL: Florida Department of Environmental Protection. http://infohouse.p2ric.org/ref/41/40276.pdf
16. Linenfelser, Brett and Lindsay Griffith. 2007. Evaluating Waterbody Assessment and Listing Processes: Integration of Monitoring and Evaluative Techniques. Water Environment Research Foundation. IWAP ISBN: 1-84339-775-7. https://www.werf.org/a/ka/Search/ResearchProfile.aspx?ReportId=04-WEM-4
17. National Research Council, 2001. Assessing the TMDL Approach to Water Quality Management. Committee to Assess the Scientific Basis of the Total Maximum Daily Load Approach to Water Pollution Reduction, Water Science and Technology Board, National Research Council. 122 pages. ISBN: 0-309-07579-3. http://onlinelibrary.wiley.com/doi/10.1029/01EO00380/pdf
18. Niemi, G.J., P. Devore, N. Detenbeck, D. Taylor, A. Lima, J. Pastor, D.J. Yount, R.J. Naiman. 1990. Overview of case studies on recovery of aquatic systems from disturbance. Environmental Management 14(5):571-587. DOI: 10.1007/BF02394710
19. SWRCB, 2004. California State Water Resources Control Board, Division of Water Quality, Functional Equivalent Document, Water Quality Control Policy for Developing California's Clean Water Act Section 303(d) List. September 2004. pp.154-208. https://www.waterboards.ca.gov/water_issues/programs/tmdl/docs/ffed_093004.pdf
20. Shabman, Leonard. A, and Smith, Eric P. 2003. Implications of applying statistically based procedures for water quality assessment. Journal of Water Resources Planning and Management 129 (4) 330-336.
21. Smith, Eric P., Keying Ye, Chris Hughes, and Leonard Shabman. 2001. Statistical Assessment of Violations of Water Quality Standards under Section 303(d) of the Clean Water Act. Environmental Science and Technology 35 (2001) 606-612.

Appendix 1
Statistical Concepts

Statistical tests take into account the uncertainty due to sample size and a user-defined error rate to determine whether a waterbody is attaining or exceeding the standard, given the information provided by the water quality samples collected. These treat the samples collected as a *random* sample of the conditions in the waterbody, rather than treating them as the entire population of the waterbody. Statistical procedures that have been identified by EPA for potential use in 303(d) listing determinations include Student's t-test for the mean, the Wilcoxon signed-rank test, lower-percentile confidence limits, the exact binomial test, the Bayesian binomial test, and the exact hypergeometric test[23],[24].

Hypothesis testing

A statistical hypothesis is an educated guess that can be tested through experimentation or observation of data. For listing determinations, we would consider two main hypotheses:

1) A waterbody is attaining a water quality criteria.
2) A waterbody is impaired for (not attaining) a water quality criteria.

A statistical hypotheses formalizes how you will assess the data about the waterbody. The result of a statistical test can indicate whether data supports the hypothesis or not, and the probability or certainty that the hypothesis is supported or not supported.

The Null hypothesis

The null hypothesis (annotated as H_0) is a hypothesis that is the assumed state. It is the hypothesis you will assume is true unless the data indicates the hypothesis is false. For instance, for listing determinations, using a null hypothesis that the waterbody is not attaining means that unless there is sufficient data significant enough to show attainment, you would assume that the waterbody should be listed.

For purposes of listing, the recommended null hypothesis is that the waterbody is not impaired. The test is whether the available data suggests the true proportion of samples in the population is less than the critical exceedance rate.[25] [26] [27] If there is enough data to reject this hypothesis, the

[23] EPA, 2002. Consolidated Assessment and Listing Methodology (CALM). Toward a Compendium of Best Practices, First Edition. United States Environmental Protection Agency. July 2002

[24] CA-SWRCB, 2004. California State Water Resources Control Board, Division of Water Quality, Functional Equivalent Document, Water Quality Control Policy for Developing California's Clean Water Act Section 303(d) List. September 2004. pp.154-208

[25] Gibbons, 2003. A Statistical Approach for Performing Water Quality Impairment Assessments. Journal of the American Water Resources Association 39(4): 841-849

[26] Linenfelser, Brett and Lindsay Griffith. 2007. Evaluating Waterbody Assessment and Listing Processes: Integration of Monitoring and Evaluative Techniques. Water Environment Research Foundation.

[27] National Research Council, 2001. Assessing the TMDL Approach to Water Quality Management. Committee to Assess the Scientific Basis of the Total Maximum Daily Load Approach to Water Pollution Reduction, Water Science and Technology Board, National Research Council.

waterbody is listed. Choosing a null hypothesis that a waterbody is “meeting water quality standards” along with a significance level of 0.05 indicates the waterbody will be assumed to attain unless there is sufficient evidence to the contrary is available.

For purposes of delisting, the recommended null hypothesis is that the waterbody is impaired. The test is whether the available data suggests the true proportion of samples in the population is higher than the critical exceedance rate. If there is enough data to reject this hypothesis, the waterbody is delisted.

The nature of a statistical hypothesis test is to determine whether there is enough evidence (data) to reject a presumption made about the population the data represents (the null hypothesis).

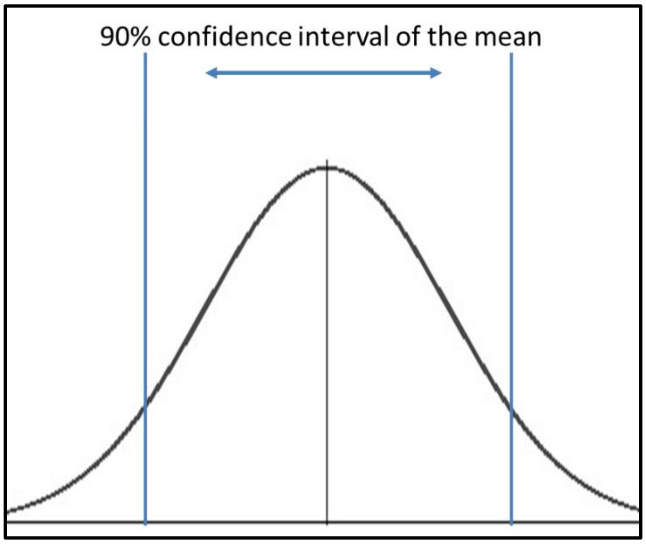
Uncertainty

Confidence

The confidence interval is the probably range over which the true population value lies. A confidence interval defines the range over which it is likely the true value of the average lies, with some quantified probability of likelihood. The size of the confidence interval depends on the range and the variability of the data in the sample.

The width of this range is determined by the variability in the sample and a pre-selected probability. Common confidence intervals are 90% and 95%. For example, when taking the average concentration of a sample, you collect multiple samples, which will not be exactly the same. To calculate the average of a population from this sample, you average the averages. The true average of the population you are measuring lies somewhere on the range of sample averages that were collected.

Figure 8. Example of a 90% confidence interval of the mean. There is 90% probability that on repeated sampling, the value of the population mean estimate falls within the interval.



Decision error rates

Type-I errors occur when a null hypothesis that is true is incorrectly rejected. Also referred to as a false-positive error rate. Type-I error is controlled by careful selection of an appropriate null hypothesis to test with the data, and the user designated significance level of the test. A higher confidence level will result in reduced type-I error. In water quality assessment, if the null hypothesis is that a standard is attained, the type-I error refers to listing a waterbody in Category 5 determination, when it actually attains the standard.

Type-II errors occur when a null hypothesis that is false is not rejected. Also referred to as a false-negative error. It is not an error in the sense that an incorrect conclusion was made about the data. Only that the default assumption was not rejected when there was an opportunity to do so. No conclusion about the assumed state of the sample is made. Rejecting the null hypothesis is a conclusion about the data. Failing to reject the null hypothesis is a statement that there was not sufficient data to conclude anything other than the default assumption. Type-II error probability depends on the sample size, significance level, and data variability. There is a tradeoff between Type I and Type II errors, as Type II error is usually the reciprocal of the Type-I error rate. Type-II error is often reduced by increasing the sample size or increasing the width of the confidence interval. In water quality assessment, a type-II error commonly refers to failure to list as Category 5 a waterbody that actually exceeds the water quality standards.

Table 8. Types of Decision Error in Hypothesis Testing [28]

Decision	Reality	
	Null is True	Null is False
Reject Null	Type-I Error (False Positive)	Correct Decision (Power)
Do not reject Null	Correct Decision (Confidence level)	Type-II Error (False Negative)

As a general, Type-I errors are considered to be more serious than Type II errors, because rejecting the null hypothesis indicates an effect was observed. Requiring strong evidence to reject the null hypothesis makes it unlikely a true null hypothesis will be rejected. Simultaneous control of both type-I and type-II error to low levels requires large sample sizes.

Effect size

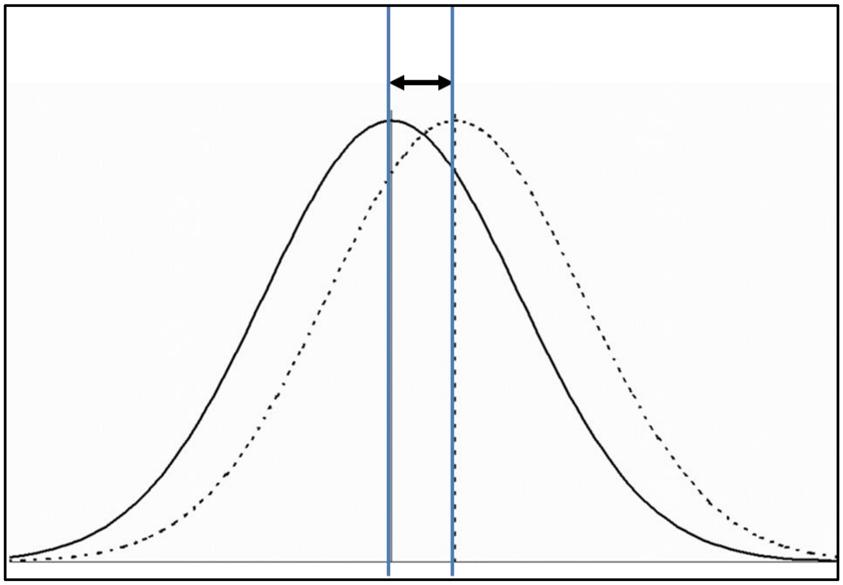
The effect size is the size of the difference between two groups or quantities where a difference between them is recognized. It sets the expectation of how different two values, distributions, or thresholds must be from each other before they are considered different from each other.

For example, the non-detect value in laboratory measurement is an effect size. It is the concentration value where the analytical precision can start to discern some small amount of an analyte from zero. In this case, the detection limit is the effect size. A sample concentration would need to be greater than this value to be distinguishable from zero.

The effect size is one of the parameters that can be adjusted to control type-II error rates.

[28] Adapted from CA-SWRCB, 2004. California State Water Resources Control Board, Division of Water Quality, Functional Equivalent Document, Water Quality Control Policy for Developing California's Clean Water Act Section 303(d) List. September 2004.

Figure 9. Effect size of the difference between the means of two samples.



Appendix 2 – Selected review of statistical listing methodologies used in other states

Table 9. Statistical methods to assess attainment of standards for toxic substances employed by other states

State	Assessment of toxic pollutant standards chronic criteria
Alaska	• Documented persistent exceedances of a criterion or criteria. • Documentation or water quality data which demonstrates designated uses are adversely affected by a pollutant condition. • *Proposed for 2018 is a binomial method with 5%-10% critical exceedance at 90% confidence or alternatively a 1-in-3-year exceedance of 2 sample critical exceedance for both acute and chronic, with a minimum sample size of 10, and chronic samples must represent a 4-day average of concentration.
Idaho	• Defines an exceedance as 2 consecutive periodic sample excursions (i.e. monthly • Critical exceedance rate of more than 1 such exceedance in a 3-year period
Washington	• Critical exceedance of 2 or more samples within a 3-year period exceed the aquatic life criteria *Currently reviewing a hypergeometric statistical method
California	• Binomial distribution with an effect size of 15% and a balanced 80% confidence level • Different null hypotheses for listing and delisting. • Listing Null Hypothesis: Actual exceedance proportion ≤ 3 percent • Listing Alternate Hypothesis: Actual exceedance proportion > 18 percent
Colorado	• The 85th percentile of the ranked samples must be less than the criteria for chronic toxic chemical standards • No single sample may exceed more than once within a 3 year period for acute toxic chemical criteria. • The 15th percentile of the ranked samples must be less than the criteria for DO, pH standards
Florida	• Binomial distribution with a ≥10% critical exceedance rate • with a minimum of an 80 percent confidence level using a

	binomial distribution • minimal sample size of 10 samples
Iowa	• Binomial distribution with a >10% critical exceedance rate and 90% confidence for conventional pollutant criteria • Raw score with >10% critical exceedance for chronic toxic substances criteria • 1-in-3-years of 2 sample critical exceedance for acute toxic substances criteria
Kansas	• Pre-screen potential impairments for > 1-in-3-year excursion of the domestic water supply nitrate criteria and acute aquatic life • Screen for a raw score 10% critical exceedance • For those sites that fail the screening (>10% critical exceedance), apply a binomial distribution of a 10% critical exceedance with a 90% confidence interval to determine impairment
Montana	• Raw score with a critical exceedance rate of >10% • at least one sample exceeds twice the acute criteria • silver has a single exceedance of the acute aquatic life standard
Nebraska	• Binomial method with >10% critical exceedance rate and 90% confidence for both acute and chronic criteria
Nevada	• Binomial method with >10% critical exceedance rate at 90% confidence • Minimum number of 3 samples needed to make an impairment determination
New Hampshire	• Raw score with 10% exceedance rate of acute or chronic criteria for impairment • Minimum sample size of 2 • Will list with <10% of samples if samples represent a very large magnitude of exceedance
North Carolina	• Binomial distribution with a critical exceedance > 10% of samples and 90% confidence • Minimal sample size >9
Pennsylvania	• “Select and apply appropriate analytical techniques” to evaluate whether samples indicate water bodies are “attaining standards at least 99% of the time.“
Texas	• Binomial method with a 10% critical exceedance rate and 80% confidence • Confidence varies by sample size • Accepts up to ~40% type-I and type -II error rate • Minimal sample size of 10

Detailed examples of states utilizing statistical assessment of toxic substances for the protection of aquatic life

California

Since 2004, California has used a binomial distribution method for their 303(d) assessment purposes. Their methodology does not require an absolute minimum number of samples but varies based on sample size. By choosing no minimum sample size, California allows smaller sample sizes to be used if the frequency of sample exceedances is large. For example, if 2 exceedances are required to list for a sample size of 16, than if two or greater exceedances occur when the total number of samples is 16 or less, than the waterbody would be placed on the 303(d) list.

For the purposes of analyzing statistical confidence and power, the null hypothesis is: water quality standards are met. The alternative hypothesis is, then, water quality standards are not met. Decisions on whether the waterbody should be listed depend on which hypothesis, the null or alternative hypothesis, is "rejected" at a certain level of confidence and power. California's null hypothesis is that the exceedance proportion of samples is less than or equal to 3 percent. The alternative hypothesis is that the actual exceedance proportion of the samples is greater than or equal to eighteen percent with a minimum effect size of fifteen percent (3% + 15% = 18%).

In other words, if the data indicates with 90% confidence that less than 3% of all possible samples for the waterbody would be above the standard, it is placed in Category 2. If the data indicates with 90% confidence that more than 18% of all possible samples for the waterbody would be above the standard, it is placed in Category 5. Conflicting evidence is considered to occur when the data indicates that the proportion of all possible samples in the waterbody expected to exceed is greater than 3% but less than 18% (equivalent to Oregon's Category 3B).

Despite the EPA CALM Guidance recommendation of a balanced error rate of 0.15, California chose to balance their error rate at 0.20. They felt that use of the higher error rate (20 percent) was substantiated because the basis for the listing would be reviewed and corroborated by subsequent analyses performed in the course of TMDL development.

Florida

The concern that the statistical analysis of water quality data will result in an inappropriate revision to existing water quality criteria has been raised across a variety of states and EPA regions and was addressed in Florida. Unlike Oregon's one in three year exceedance frequency criteria, Florida's currently applicable water quality standards state that, "unless otherwise stated, all criteria express the maximum not be exceeded at any time." Despite this not to be exceeded rule, the primary feature of the Florida Impaired Waters Rule (IWR) is the use of a statistical test based on the binomial distribution to evaluate data sets of water quality parameter measurements prior to relying on such data sets in listing a waterbody as "impaired."

Florida uses a 10% probability value which is a sample exceedance rate for the assessment data, not an inherent allowable rate of criteria exceedance in the ambient water. In 2005, EPA determined the probability value was a new or revised water quality standard as a change to the frequency component of criteria. In their 2007 review, EPA modified their determination based on additional information submitted by the Florid Department of Environmental Protection. EPA concluded that the probability value is a data reliability component of the Impaired Waters Rule rather than a modification to the frequency component of the criteria. EPA's current interpretation is that the purpose of the 10% probability value is to exclude data that are likely to be unrepresentative of actual ambient water conditions.

For their impaired waters determination, Florida identifies both a "planning" and a "verified" list. It is customary for scientists and decision-makers to look for a high degree of confidence in order to reject a null hypothesis. Any statistical conclusion that has a confidence level of less than 90 percent is generally considered not acceptable by most statisticians[29]. Consequently, Florida requires a 90% confidence limit for the "verified" list (i.e. 90% confident the waterbody is impaired) and an 80% confidence limit for the planning list. The probability value, however, remains the same.

North Carolina

EPA's review of North Carolina's 2016 303(d) submittal concluded that the state's assessment approach was acceptable for most listing decisions, but not all listing decisions. The EPA determined that North Carolina DEQ's methodology did not reasonably assess toxic or non-conventional pollutants consistent with the State's applicable and EPA-approved water quality standards (WQS). In addition, the EPA determined that the state's methodology did not contain defensible, statistically-sound delisting procedures for most numeric WQS. This led to a failure to demonstrate good cause to delist impaired waters.

Similar to Florida, North Carolina's water quality standards for toxics are specified as "maximum permissible levels." Because the State's standards do not define the conditions of toxicity (acceptable duration and frequency), one basic interpretation of the water quality standards could be that no exceedances are permissible in the waters of the state; i.e., one sample value over the applicable criterion is cause for listing the water as impaired. In 2016, North Carolina, used the binomial distribution to assess its waters for toxics impairment by using a 10% probability value with greater than or equal to a 90% statistical confidence level for sample sizes greater than nine.

North Carolina justified it use of the binomial distribution for assessment of toxics by arguing that the one in three year exceedance frequency was not a valid frequency for the assessment of chronic standards. North Carolina suggested that a modernization of water quality assessment of chronic standards for toxics is appropriate since the one in three year frequency was based on outdated studies that reflected acute events. The State's justification, however, did not address

[29] Lin, P., M. Duane, and X.F. Niu. 2000. Nonparametric procedure for listing and delisting impaired waters based on criterion exceedances. Task I, Contract Number LAB015 Tallahassee, FL: Florida Department of Environmental Protection

how a ten percent exceedance rate with a 90% confidence level supported attainment of water quality standards or demonstrate protection of aquatic life.

In their approval letter, EPA stated that North Carolina DEQ was not required to use the EPA-recommended one-in-three year method; however, North Carolina had not provided a scientifically defensible rationale to support their listing methodology for toxics. In the state's Section 303(d) list submittal of April 1, 2016, North Carolina provided a "White paper" entitled Water Quality Assessment Methods for Toxics to provide "a scientific basis, rationale and justification for not relying on exceptionally small datasets for making a 303(d) listing decision." EPA concluded this document provided a "Retrospection of the '>1-in-3' Assessment Method," but it did not provide a rationale to support a ten percent exceedance rate at a given confidence level.

Other States

Other states that have adopted, or are in the process of adopting, the binomial distribution method include: Alaska, Texas, Nebraska, and Kansas (Table 4). Kansas lists waters based on assessment of exceedances of the acute standard first, then uses the 10% raw score for chronic data as a second screen. If an impairment is detected using the raw score, then the binomial test is used to confirm the data indicates the standard is exceeded. Washington has proposed to use a different statistical procedure based on the hypergeometric mean, which is similar to the binomial test, but assumes there is a finite population of samples. Texas uses a binomial test with variable confidence levels for both toxics and conventional parameters in their assessment.[30]

[30] Texas Commission on Environmental Quality, Surface Water Quality Monitoring Program, 2014 Guidance for Assessing and Reporting Surface Water Quality in Texas (June, 2015); pp 2-18.

State Assessment Methodology Documents

1. Alaska Water Quality Monitoring and Assessment Strategy. May, 2015. https://dec.alaska.gov/water/wqsar/monitoring/DEC_monitoring_strategy_final_2005.pdf
2. California Water Boards Water Quality Control Policy for Developing California's Clean Water Act Section 303(d) List. 2004. http://www.waterboards.ca.gov/water_issues/programs/tmdl/docs/ffed_303d_listingpolicy093004.pdf
3. Colorado Section 303(d) Listing Methodology. 2018 Listing Cycle. https://www.colorado.gov/pacific/sites/default/files/303d_LM_2018.pdf
4. Florida DEP Identification of Impaired Waters, 2012. https://www.flrules.org/gateway/notice_Files.asp?ID=17118286
5. Idaho Waterbody Assessment Guidance, 2016. http://www.deq.idaho.gov/water-quality/surface-water/monitoring-assessment/
6. Iowa Methodology for Iowa's 2016 Water Quality Assessment, Listing, and Reporting Pursuant to Sections 305(b) and 303(d) of the Federal Clean Water Act. March 28, 2017. http://www.iowadnr.gov/Portals/idnr/uploads/watermonitoring/impairedwaters/2016%20Iowa%20Draft%20Methodology.pdf?ver=2017-04-11-134154-320
7. Kansas Methodology for the Evaluation And Development of the 2016 Section 303(D) List Of Impaired Water Bodies For Kansas. February 18, 2016. http://www.kdheks.gov/tmdl/2016/2016_303_d_Methodology_Feb-18-2016_Final.pdf
8. Montana Department of Environmental Quality Metals Assessment Method, 2012. https://deq.mt.gov/Portals/112/Water/WQPB/.../Metals_Assessment_Method.pdf
9. Nebraska Methodologies for Waterbody Assessments and Development of the 2016 Integrated Report for Nebraska. July, 2015. http://deq.ne.gov/Publica.nsf/xsp/.ibmmodres/domino/OpenAttachment/Publica.nsf/53AA14CA60E0CE1486257FA1006655A8/Attach/2016%20IR%20Assessment%20Methodology.pdf
10. Nevada 2014 Water Quality Integrated Report. http://ndep.nv.gov/bwqp/file/IR2014_Report.pdf
11. New Hampshire State of New Hampshire Draft - 2016 Section 305(b) and 303(d) Consolidate Assessment and Listing Methodology. https://www.des.nh.gov/organization/divisions/water/wmb/swqa/documents/calm.pdf
12. North Carolina 2016 303(d) Listing Methodology. https://ncdenr.s3.amazonaws.com/s3fs-public/Water%20Quality/Planning/TMDL/303d/2016/2016%20Listing%20Methodology%20approved%20by%20EMC%20May%202015.pdf
13. Pennsylvania 2016 Draft Integrated Water Quality Monitoring and Assessment Report. http://www.elibrary.dep.state.pa.us/dsweb/Get/Document-113834/2016_Draft_Pennsylvania_Integrated_Water_Quality_Monitoring_and_Assessment_Report_Updated_07-28-2016.pdf
14. Tennessee 2014 305(b) Report: The Status of Water Quality in Tennessee. http://www.tennessee.gov/assets/entities/environment/attachments/wr_wq_report-305b-2014.pdf
15. Texas 2014 Guidance for Assessing and Reporting Surface Water Quality in Texas. June, 2015.

https://www.tceq.texas.gov/assets/public/waterquality/swqm/assess/14txir/2014_guidance.pdf

16. Utah’s 303(d) Assessment Methodology. May, 2016. https://deq.utah.gov/ProgramsServices/programs/water/monitoring-reporting/assessment/2015MarDraft303dMeth.htm
17. Washington Department of Ecology Water Quality Policy 1-11, 2012 http://www.ecy.wa.gov/programs/wq/303d/policy1-11.html

Appendix 3

Calculation of critical values for applying the exact binomial test

Summary

States are required to establish a list of impaired waterbodies under Section 303(d) of the Clean Water Act. The Oregon DEQ is proposing to evaluate attainment of numeric water quality criteria for aquatic life toxics and conventional pollutants using the exact binomial hypothesis test. This will enable Oregon to determine if data indicates waterbodies are exceeding criteria with a reasonable level of confidence.DEQ is not proposing to apply the binomial statistical test for assessment of acute standards or human health criteria.

The State of Oregon currently assesses compliance with numeric biologically-based aquatic life criteria for impairment using a simple 2 sample threshold for toxic substances. If any two observed samples within a waterbody exceed the criteria value, it is grounds for listing a waterbody as impaired (Category 5) on the 303(d) list. This matched the >1-in-3-year frequency of exceedance used for acute toxic assessment. Conventional pollutants are assessed for compliance using the "10% rule," a raw score method that considers a waterbody to be impaired if more than 10% of the observed samples are above the criterion threshold.

DEQ has received significant input from stakeholders that the current assessment method for toxic substances is likely to overestimate the number of impaired waters, particularly in cases where sample size is large and the vast majority of samples are below the criteria thresholds. Additionally, there has not been a clear set of requirements for removal of a waterbody from the 303(d) list when improvements in water quality indicate the waterbody is currently attaining the criteria.

DEQ is proposing to update the listing and delisting methodology for assessing toxic substances and conventional pollutants using a one-sample test on binomial proportions, the exact binomial test. The binomial test allows for a statistically valid assessment of the proportion of samples in a waterbody that would be expected to exceed the criterion while accounting for uncertainty in the assessment procedure.

The proposed assessment methodology attempts to balance desired data requirements with the practical realities affecting the availability of information and the strength of the available evidence. In order to address the propensity for small data sets, DEQ will continue to apply the >1-in-3-year critical exceedance rate to all data sets less than a minimum number of samples.

Background

The exact binomial test directly limits the type-I error probability (the probability of incorrectly rejecting a null hypothesis) at or below a chosen nominal significance level (α). Using the binomial distribution, the critical value (the number of samples that exceed the threshold set by the criterion) can be calculated for a given sample size such that α is $\geq$ the desired confidence level. The type-II error rate is indirectly determined by sample size. However, balancing of α and β error rates can be accomplished by selection of complimentary critical rates for simultaneous evaluation of attainment and impairment in the same data set.

This attachment details the methodology and probability equations used to derive tables of critical values and both α and β rates for each proposed formulation of the test.

The binomial test determines the likelihood that the number of values exceeding the criterion threshold, *k*, observed in a sample of the population indicates that the true exceedance rate in the population, *r*, would be greater than the regulatory critical exceedance rate, *p*.

A water body is determined to be impaired if the number of excursions, *k*, in *n* samples equals or exceeds the critical value for impairment, k_i. A water body is determined to be attaining if $k \leq k_i$ for a given *n*. An alternative critical value for attainment, k_a, can be used to determine attainment if the desired error probability is different from that for impairment. In this case, a waterbody would be considered attaining if $k \leq k_a$.

The value of k_i and k_a are determined iteratively as the largest number of exceedances, respectively, such that 1-α is within the desired confidence interval for a sample of size n, with a regulatory critical exceedance rate threshold, *p*. EPA recommends an effect size of 15% for determining the difference between attaining and impaired waters. Therefore, the difference between the acceptable critical exceedance proportion, p_1, and the unacceptable critical exceedance proportion, p_2, should be such that $p_2 – p_1 = 0.15$.

Note that β is not considered in choosing the value of either k_i or k_a. The following procedures for listing and delisting are based on keeping α at or below a pre-determined limit but do not directly limit β. In this proposal, p_1 is used to calculate α, and p_2 is used to calculate β. However, only α, and consequently p_1, has an effect on the value of k_i or k_a. Tables for listing and delisting that show the value of k for a range of samples are produced below.

Alternately, selection of *k* could be based on optimization of the minimum achievable α and β, without requiring either to be held below a maximum level. This process is known as error balancing.

Generalized Listing Procedure

The listing procedure is based on a default assumption that the true exceedance rate, *r*, is less than or equal to the regulatory exceedance rate, p_1.

The tested one-sided hypotheses are:
H_O: $r \leq p_1$
H_A: $r > p_1$

Then calculate α from the right tail probability of the cumulative binomial distribution:

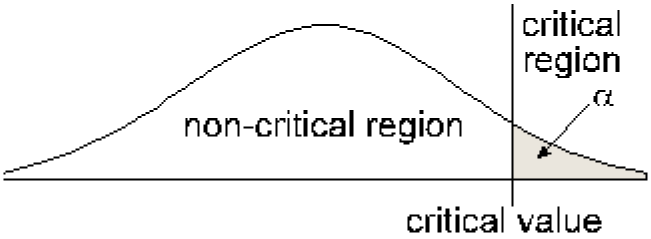
Where, n = the number of samples,
k_i = the critical value of the minimum number of sample excursions needed to place a water on the section 303(d) list,
p_1 = regulatory critical exceedance rate, and
p_2 = unacceptable critical exceedance rate.

BINOMDIST( ) is an Excel® software function that returns cumulative left tail binomial probabilities.

For sample sizes of n, the minimum number of measured exceedances is established where α < 0.1, and where |α - β| is minimized.

α = Excel® Function BINOMDIST(n-k_i, n, 1 – p_1, TRUE)
β = Excel® Function BINOMDIST(k_i-1, n, p_2, TRUE)

The minimal sample size for listing and delisting was selected based on the sample size needed to achieve sufficient power with a minimum of 2 excursions, where 1- β > 0.90.
The value of k_i=2 for n=2, and increased by 1, where 1-α < 0.90

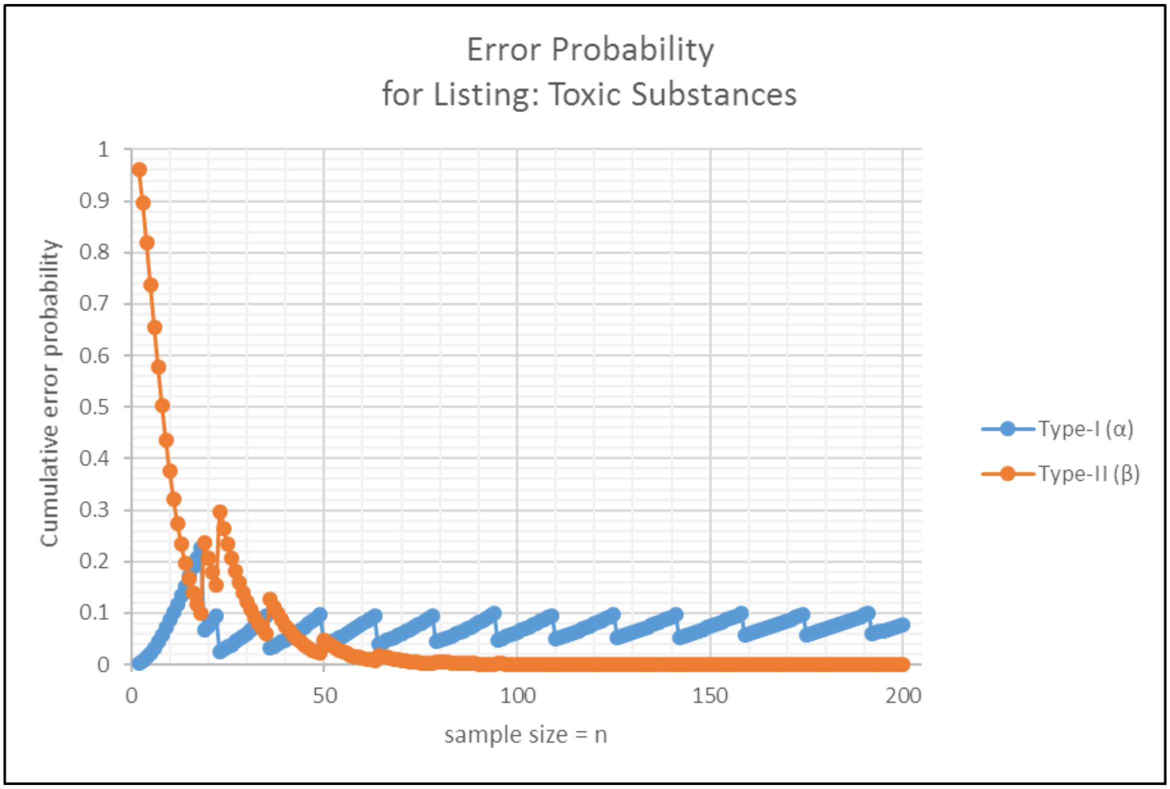


Critical Values for Listing Chronic Toxic Substances

Null Hypothesis: Actual exceedance proportion is ≤5%
Alternate hypothesis: Actual exceedance proportion is >5%
Minimum confidence level is 90%

The minimum sample size necessary to attain 90% confidence in H_A with a minimum of 2 excursions is 18, but this critical rate is extended to smaller samples sizes.

<table>
<tr><th colspan="2">Minimum number of sample excursions required to list as impaired for toxic substances</th></tr>
<tr><th>Sample Size</th><th>Minimum number of excursions</th></tr>
<tr><td>2-18</td><td>2*</td></tr>
<tr><td>19-22</td><td>3</td></tr>
<tr><td>23-35</td><td>4</td></tr>
<tr><td>36-49</td><td>5</td></tr>
<tr><td>50-63</td><td>6</td></tr>
<tr><td>64-78</td><td>7</td></tr>
<tr><td>79-92</td><td>8</td></tr>
<tr><td>93-109</td><td>9</td></tr>
<tr><td>110-125</td><td>10</td></tr>
<tr><td>126-141</td><td>11</td></tr>
<tr><td>142-158</td><td>12</td></tr>
<tr><td>159-171</td><td>13</td></tr>
<tr><td>179-191</td><td>14</td></tr>
<tr><td>192-200</td><td>15</td></tr>
<tr><td colspan="2">* The use of 2 excursions to list is extended for sample sizes <10 in order to achieve >90% confidence in H_A for small sample sizes.</td></tr>
</table>

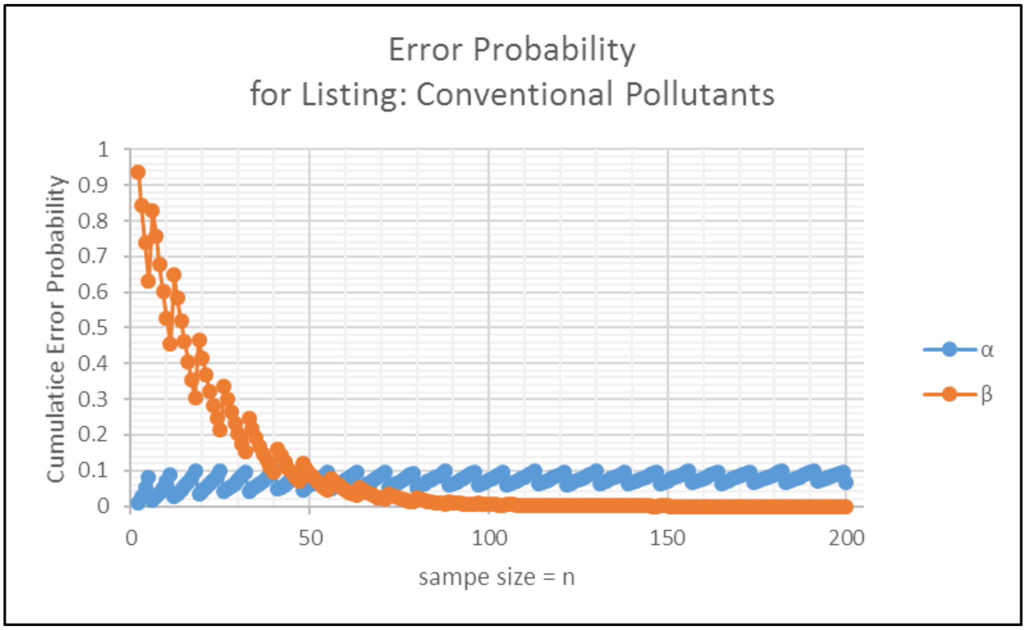

Error Probability
for Listing: Toxic Substances
Cumulative error probability
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
50
100
150
200
sample size = n
Type-I (α)
Type-II (β)

Critical Values for Listing Conventional Pollutants

Null Hypothesis: Actual exceedance proportion is ≤10%
Alternate hypothesis: Actual exceedance proportion is >10%
Minimum confidence level is 90%

The minimum sample size necessary to attain 90% confidence in H_A with a minimum of 2 excursions is 18, but this critical rate is extended to smaller samples sizes.

Minimum number of sample excursions required to list as impaired for conventional pollutants	
Sample Size	**List if ≥ :**
5 - 11	2*
12-18	4
19-25	5
26-32	6
33-40	7
41-47	8
48-55	9
56-63	10
64-71	11
72-79	12
80-88	13
89-96	14
97-104	15
* The use of 2 excursions to list is extended for sample sizes <11 in order to achieve >90% confidence in H_A for small sample sizes.	


Error Probability
for Listing: Conventional Pollutants
Cumulatice Error Probability
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
50
100
150
200
sampe size = n
α
β

Generalized Delisting Procedure

For finding a waterbody is attaining, we want to know the critical value for samples that indicate the true proportion in the population, *r*, is greater than the regulatory critical exceedance rate, p_1.

H_O: $r > p_1$
H_A: $r \leq p_1$

Then calculate α from the left tail probability of the cumulative binomial distribution:

Where n = the number of samples,
K_a = maximum number of measured exceedances to determine a water is attaining, or should be removed from the 303(d) list,

p_1 = unacceptable exceedance proportion, and
p_2 = acceptable exceedance proportion.

BINOMDIST( ) is an Excel software function that returns cumulative left tail binomial probabilities.

For sample sizes of n, the minimum number of measured exceedances is established where $\alpha < 0.10$ and $1-\alpha > 0.90$, and where $|\alpha - \beta|$ is minimized.

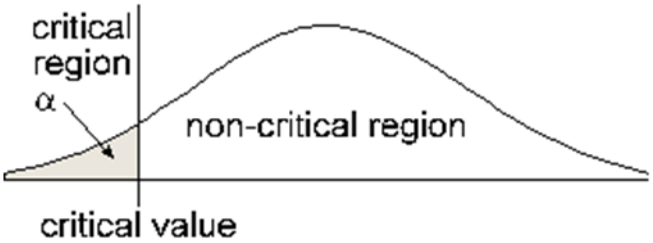
α = 1 – Excel® Function BINOMDIST (k_a-1, n, p_1, TRUE)

β = 1 – Excel® Function BINOMDIST (n-k_a-1, n, 1-p_2, TRUE)

The minimal sample size for delisting was selected based on the sample size needed to achieve sufficient power with a minimum of 1 excursion, where $1- \beta > 0.90$.

The value of k_a=1 for n=10, and increased by 1, where $1-\alpha < 0.90$

(See attached Spreadsheet)

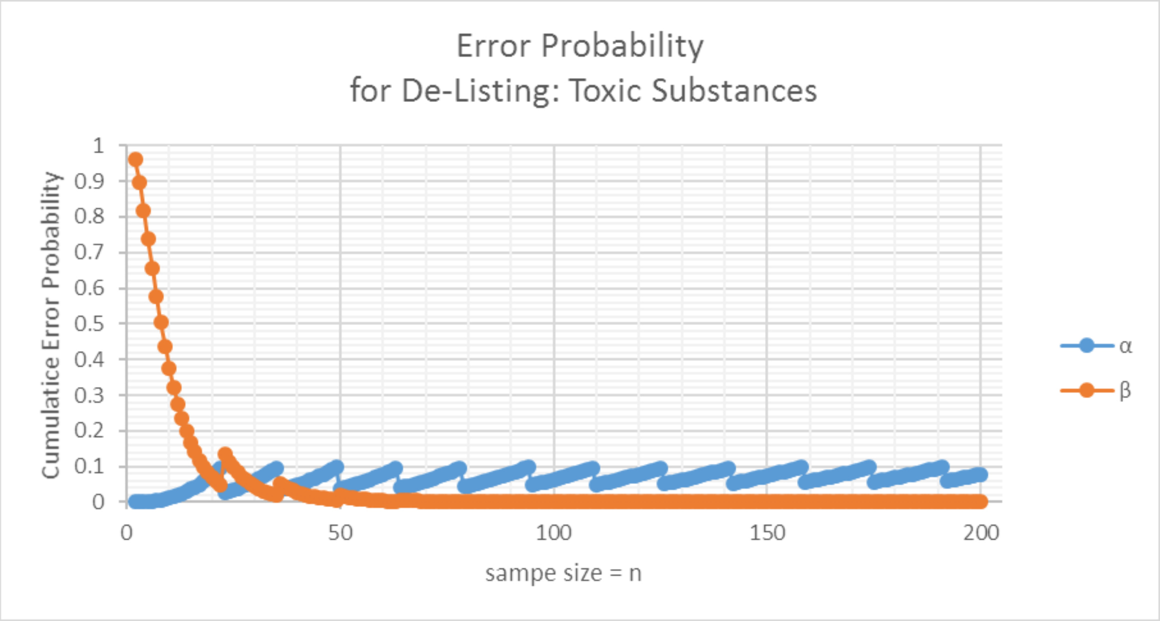


Critical Values for Delisting Chronic Toxic Substances

Null Hypothesis: Actual exceedance proportion is >5%
Alternate hypothesis: Actual exceedance proportion is ≤5%
Minimum confidence level is 90%

The minimum sample size necessary to achieve 90% confidence for H_A with 1 exceedance is 18.

Maximum number of sample excursions to delist as impaired for toxic substances	
Sample Size	**Delist if ≤ :**
18-22	1
23-35	2
36-49	3
50-63	4
64-78	5
79-94	6
95-109	7
110-125	8
126-141	9
142-158	10
159-174	11
175-191	12
192-200	13

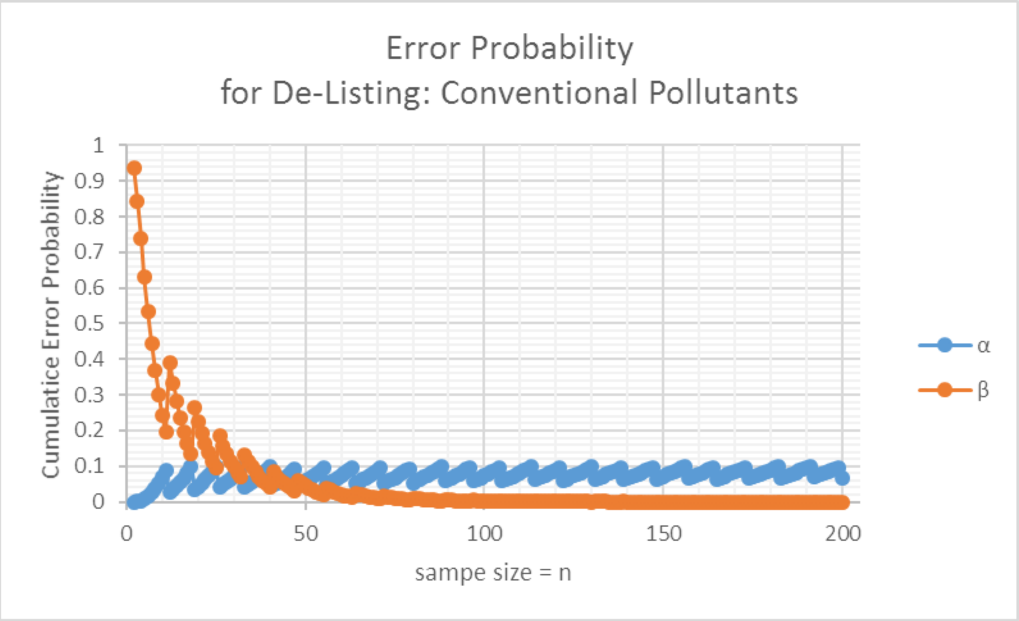

Error Probability
for De-Listing: Toxic Substances
Cumulatice Error Probability
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
50
100
150
200
sampe size = n
α
β

Critical Values for Delisting Conventional Pollutants

Null Hypothesis: Actual exceedance proportion is >10%
Alternate hypothesis: Actual exceedance proportion is ≤10%
Minimum confidence level is 90%

The minimum sample size necessary to achieve 90% confidence for H_A with 1 exceedance is 15.

Maximum number of sample excursions to delist as impaired for toxic substances	
Sample Size	**Delist if ≤ :**
15	1
16-18	2
19-25	3
26-32	4
33-40	5
41-47	6
48-55	7
56-63	8
64-71	9
72-79	10
80-88	11
89-96	12
97-104	13


Error Probability
for De-Listing: Conventional Pollutants
Cumulatice Error Probability
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
50
100
150
200
sampe size = n
α
β

Appendix 4

External Peer Review

Section 1

Introduction

This section contains the responses to all comments received from the peer review panel on the proposed application of the exact binomial test for assessment of chronic aquatic life criteria for toxic pollutants, and for conventional pollutants.

DEQ compiled a review of statistical methods used by other states, and supported by EPA guidance, in a whitepaper drafted in October 2017. The peer review panel was convened in December of 2017. DEQ solicited potential panel members from DEQ staff, EPA staff, and stakeholders involved in a stakeholder workgroup for improvement of the Integrated Report methodology.

The panelists completed review of the methodology on January 29, 2018. A revised draft of the whitepaper will be provided to DEQ's stakeholder workgroup to identify any and discuss any resultant policy issues. Following any additional policy input, the resultant draft assessment methodology including the method based on this work will be made available for public review and comment in March 2018.

Panel members are listed in Section 1.2 and are identified by number. A summary of all comments submitted and DEQ's response is presented in Section 2. Comments that addressed the same issue were grouped and a common response was given to address the comment. Unique comments were answered individually. The original panel response forms are appended to the end of this document.

Section 1.2 List of Panelists

1. Dr. Gerrad Jones, PhD.
   Assistant Professor of Biological and Ecological Engineering
   Oregon State University

2. Dr. Douglas McLaughlin, PhD.
   Principal Research Scientist
   National Council for Air and Stream Improvements (NCASI)

3. Jason Law, M.A.
   Statistician
   City of Portland Bureau of Environmental Services

4. Dr. Jon Harcum, PhD.
   Principal engineer, hydrologist / Engineering Lecturer
   Tetra Tech, Inc. / Clemson University

5. Dr. Yangdong Pan, PhD.
   Professor of Environmental Science & Management
   Portland State University

6. Patrick Moran, M.S.
   Aquatic Toxicologist
   U.S. Geological Survey

Section 1.3 Peer review response overview

Summary

All panelists generally agreed DEQ's proposed application of the binomial test is appropriate and defensible. They also considered the use of the binomial test an improvement over current practice. DEQ should provide more explanation for the proposed critical exceedance rates and confidence levels. None of the panelists stated there were any errors in DEQ's method for calculating the critical values. Calculations for the critical number of excursions to use for listing and delisting for a given sample size are correct.

Several panelists identified sources of potential uncertainty due to the design of DEQ's monitoring program and data sources for the Integrated Report. The Integrated Report does not have a purpose-made sampling design. The report is required to consider all data of sufficient quality that is publicly available or submitted by third parties. Panelists identified two alternative statistical methods to address these sources of uncertainty. These are a Bayesian inference component to the binomial test, or general linear mixed model (GLMM) hypothesis test.

1. Is DEQ's proposed use of the exact binomial statistical test valid and defensible for assessment of:

- **a. chronic aquatic life toxics criteria?**
- **b. conventional pollutant criteria?**

All panelists acknowledged that DEQ's proposed application of the exact binomial test for the purposes of assessing chronic criteria for toxics substances and conventional pollutant criteria for the protection of aquatic life is valid and defensible. Several commenters noted that the adoption of this listing and delisting methodology is a marked improvement over DEQ's current methodology. One panelist suggested that DEQ consider adopting the binomial test for the assessment of acute toxic substances criteria using the same parameters as chronic toxic substances.

One panelist was concerned with sources of uncertainty that are not well controlled given the lack of a purpose-built monitoring design for the assessment. Specifically, the concern is the effect of pooling data from multiple monitoring locations within an assessment unit to evaluate attainment of the entire assessment unit.

These concerns are not specific to the application of the binomial approach. They reflect constraints imposed on the assessment program by the requirement to evaluate all publicly available data from federal agencies and additional data submitted by external parties. The Integrated Report does not have a purpose-made sampling design, and is legally required to consider data that may have been originally collected for other purposes. Collection of these data is often for other purposes and do not necessarily match the ideal requirements for a monitoring program designed specifically for identifying water quality impairments on a statewide scale.

Two panelists recommended that DEQ investigate alternative methods to address this uncertainty. They noted that the observational nature of the data structure of the assessment increases uncertainty in the degree of representativeness of samples from within a waterbody. This potentially leads to instances where data within an assessment unit does not meet all assumptions for the binomial test. Two alternative statistical tests were suggested which could account for these situations. These were general linear mixed models (GLMM) and Bayesian inference with the binomial test.

To date, there are no examples of other states applying either of these approaches to analysis of assessment data. There is also no relevant guidance from the EPA. DEQ's current proposal is in line with the procedures used in eleven states that currently apply the binomial test for 303(d) assessment. Conducting hypothesis tests using these methods would add complexity to the assessment method, and the process would be difficult to communicate to stakeholders, reducing transparency in listing decisions. At this time, DEQ is not prepared to develop new protocols and provide the needed justification to be able to propose either method potential adoption. EPA has not issued any guidance on the use of these methods, and to our knowledge, no other states have sought approval to apply them in their assessments.

2. Please comment on the selection of a confidence level of 90% for application of the binomial test for assessment purposes.

Four panelists considered a 90% confidence level to be appropriate and defensible for this application of the binomial test. One panelist recommended an 80% confidence level, citing the variability in environmental data and another panelist identified confidence levels from the range of 80%- 90% would be appropriate.

Of the eleven states applying a binomial test for assessing criteria, nine use a 90% confidence level. California and Texas use an 80% confidence level. EPA recommends a 90% confidence level in their listing guidance[31]. Selection of a confidence level within the accepted range of

[31] {EPA`;, 2002 #110}

80%-90% is partly a matter of policy and risk tolerance. Reducing the confidence level from 90% to 80% would reduce the threshold for listing waterbodies, increasing the frequency of listing impaired waters and potentially increasing the false-positive (Type-I) error rate of identifying impairments. States that use an 80% confidence level offset this lower certainty by balancing error rates. Balancing sets a greater error probability for listing decisions (higher type-I error) which reduces confidence, in order to reduce the error probability for attaining decisions (lower type-II error). By setting these at an equal rate, the chance of making an error in listing or attaining decisions is equalized.

3. Please comment on the validity of the proposed null hypotheses and critical exceedance rates:

Where r = the true proportion of sample excursions in the waterbody, and

p_1 = the acceptable regulatory critical exceedance rate, and

p_2 = the unacceptable regulatory exceedance rate, for a desired effect size of 15%.

	Toxic Pollutants (chronic criteria)	**Conventional Pollutants**
Listing	H_O: $r \leq p_1 = 0.05$, H_A: $r > p_2 = 0.20$	H_O: $r \leq p_1 = 0.10$, H_A: $r > p_2 = 0.25$
Delisting	H_O: $r > p_1 = 0.05$ H_A: $r \leq \mathrm{p}_2 = 0.20$	H_O: $r > p_1 = 0.10$ H_A: $r \leq \mathrm{p}_2 = 0.25$

Three panelists considered the critical exceedance rates for the hypothesis test to be adequate and defensible. Two panelists recommended providing more justification for the selection of the critical exceedance rates. DEQ based selection of the critical exceedance rates on EPA's (2002) guidance. EPA intended the 5% (0.05) exceedance rate for toxic substances, and the 10% (0.10) exceedance rate for conventional pollutants to reflect the desired frequency component of nationally recommended water quality standards. While DEQ may have flexibility to select exceedance rates that are more stringent, it likely has limited ability to make the exceedance rates more lenient without significant documentation and justification that alternative exceedance rates protect beneficial uses.

Two panelists noted that the formulation of the null hypothesis and alternate hypotheses were non-standard. Two panelists recommended that DEQ reformulate the null hypothesis in a standard fashion in terms of p_1 only, instead of a different p_2 for the alternative hypothesis (H_A).

DEQ intended the different exceedance rates reflecting H_O and H_A to reflect a 15% effect size as recommended by EPA. California uses a similar formulation of hypotheses for the binomial test. The observed proportion of exceedances in a sample (r/n) has a strong effect on the Type II error probability. The difference between the observed proportion of excursions in the sample and the criterion value (p_1=0.10) can be considered an effect size measure. For a specified α-level, whenever the lower bound on the estimate of r is > 0.10, we would reject H_0 and conclude that the sample is evidence that the proportion of excursions in the waterbody are over the threshold. When the specified α-level is 0.10, the lower bound of the 90% confidence interval is less than 10% until the proportion of excursions is at least 15%.

Three panelists supported use of the 15% effect size - citing it as consistent with methods in other states. They did not comment on DEQ's incorporation of the effect size as the critical exceedance rate of the alternative hypothesis, p_2. A different panelist recommended reducing p_2, the unacceptable exceedance rate or listing exceedance rate, for chronic toxic substances criteria from 0.20 to 0.15, and for conventional pollutants from 0.25 to 0.20. One panelist recommended that DEQ more clearly emphasize that the 15% effect size is used for error estimation but explicitly note that it does not affect the hypothesis test.

In effect, DEQ's hypothesis formulation provides the same outcome as the standard hypothesis formulation for listing and delisting. The p_1 values are used to reject or accept H_O. The p_2 values are used to calculate the type-II error probability, but do not affect rejection of $H_{O.}$ DEQ will emphasize that p_2 does not affect the calculation of critical values for listing or delisting.

4a. In your professional opinion, are the selected hypothesis tests and type-I error rates sufficient to minimize environmental risk from making errors in conclusions to list or delist waterbodies?

All panelists agree that the type-I error rate (α) of 10% proposed by DEQ was appropriate and suitable for addressing errors in the listing process. One panelist indicated the adoption of a method that directly limits type-I errors is a significant improvement over DEQ's previous listing methodology.

4b. In your professional opinion, does the type-II error rate for the selected critical values for listing and delisting require balancing?

Three panelists indicated that selection of the type-I and type-II error rates were up to the risk manager. Adjustment of the error rates should be based on DEQ's tolerance for making different types of decision errors. One panelist indicated that the Type II error rate appeared reasonable. None of the panelists indicated that an error balancing approach to simultaneously optimize the type-I and type-II error was necessary. One panelist considered type-II errors as more environmentally costly because they fail to identify impaired waterbodies.

EPA guidance discusses the error balancing concept, but it is not required. Error balancing reduces the type-II error probability at the expense of increased type-I errors. DEQ did not include an error balancing approach in this proposal.

4c. If so, suggest alternative methods for calculating critical values while balancing the Type I and Type II error rates.

Two panelists suggested that adding Bayesian inference to the binomial test could improve accuracy of listings. Both of these panelists also suggested that collecting more data is a reliable way to increase certainty in the assessment. One of the panelists recommended Bayesian inference as a better solution to the problem of low sample size confidence than error balancing.

DEQ does not always have the resources to collect additional data before being required to make listing decisions on waterbodies. This is especially the case where it is required to consider third party datasets in the assessment– where control of the sampling design and follow-up monitoring is not possible.

To apply a Bayesian approach, DEQ would adjust the probability that a waterbody is impaired using prior information. For instance, it could assign a higher prior probability to a waterbody being assessed where there are existing listings from adjacent waterbodies. DEQ is not aware of other states that apply a Bayesian inference method in their listing methodology. Bayesian inference relies on subjective assessment of prior probability and complicated calculations. Prior probabilities would need to be prepared by staff for each assessment unit. DEQ is concerned that this alternative would be difficult to communicate to the public and staff, complicate the assessment process, and give the appearance of less transparency and objectivity.

Section 2 Detailed Summary of Specific Comments and Responses

Section 2.1 Key to Detail Comment Summaries

Panelist comments and DEQ's responses are compiled in tabular form in Section 2.2.

Column 1: Comment number.
Consists of two numbers separated by a period. The first number corresponds to the review charge questions. The second number identifies each unique comment topic for that question.

Column 2 Panelist number.
A number identifying the panelist from the list of panelists (Section 1.2).

Column 2 Summary of comment
This column contains summaries of the peer review responses. When multiple commenters are listed, they each provided very similar comments in that area captured by the summary.

Column 3 Response to comment
This column has a short response from DEQ on the comment.

Column 4 Revision
This column states whether the methodology and/or whitepaper were revised based on the comment.

Section 2.2 Detailed Summary of Comments

Question 1

1. Is DEQ’s proposed use of the exact binomial statistical test valid and defensible for assessment of:
    a. chronic aquatic life toxics criteria?
    b. conventional pollutant criteria?

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
1.1	1,2,3,4,5,6	DEQ’s application of the exact binomial method is appropriate and defensible.	The binomial method is a standard method to assess water quality, sanctioned by the U.S. EPA and in use or under consideration in a similar form in at least eleven other states.	None required.
1.2	2,3,5	The proposed assessment method provides an improved basis for assessment decisions over the current practice.	Comment noted.	None required.
1.3	1	For the lowest number of critical values (exceedances 2), for sample size 2-18 there is little statistical power. DEQ acknowledges this and the usage of the binomial test in this range is well justified and appropriate.	DEQ has limited ability to delay making listing decisions when there is evidence of impairment. For this reason, it has retained the status quo to list based on two or more sample excursions for sample sizes less than 18. Applying the binomial method to reduce error and provide greater certainty in listing decisions based on larger data sets will provide incentive for the submission of larger data sets by many stakeholders. Larger data sets would allow for more accurate characterization of the	None required.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
			condition of waters within the state.	
1.4	6	The proposed approach to apply different statistical methodologies depending upon sample size is consistent with basic statistical principles.	Comment acknowledged.	None required.
1.5	2	DEQ should expand use of the binomial approach to assess acute aquatic life toxics criteria, consistent with EPA guidelines.	DEQ's main justification for applying the >1-in-3-year critical exceedance rate to acute toxics was the assumption that the sampling duration represents a reliable 1-hour average of pollutant concentration. This matches the duration component of the acute aquatic life criteria. DEQ acknowledges the panelist's observations that sampling variation, analytical error, and spatial and temporal variability apply to evaluation acute toxics criteria as well as chronic. The EPA consolidated assessment and listing methodology allows for adoption of the 5% critical exceedance rates to the assessment of acute toxics criteria as well as to chronic toxics criteria.	Refer to staff and advisory committee for consideration.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
			Therefore, DEQ has the option of adopting a 5% critical exceedance rate for acute toxic substances criteria as well as chronic criteria. As most toxic substance listings to date have been for violations of chronic criteria, the potential effect on listings is unknown.	
1.6	2,5	The validity of a statistical hypothesis testing largely depends on how well the population of interest is adequately sampled. DEQ must keep in mind the assumption that a set of individual measurements represents a random sample that can be used to make inferences about true condition of a waterbody relative to a numeric criterion with respect to time and space. These assumptions are important when developing monitoring programs or selecting data for assessment purposes. Decoupling the monitoring program from hypothesis testing may potentially give the public an impression that the listing or delisting is scientifically defensible simply because a well-established statistical method is used.	DEQ's monitoring programs are designed to provide accurate, representative samples of water quality within waters of the state. However, the Clean Water Act requires states to consider all readily available data that meets reasonable quality assurance requirements from other entities, including government agencies and the public, when determining assessment conclusions for the integrated report. As such, the assessment methodology cannot count on the same level of control over sampling design as if it were a completely designed and controlled experiment.	None required.
1.7	5	The data "pooling" that DEQ is considering performing violates the independence and identically distributed assumption of the exact binomial test. The method does not address the unbalanced data sets that will be	DEQ uses a targeted monitoring design for general water quality (ambient) and toxic substances. Previously, these monitoring stations were assessed individually. DEQ's change to assess data at the level of fixed assessment	Revisions to assessment unit white paper.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
		regularly encountered when performing these binomial tests.	units may include data from more than one monitoring station. These fixed assessment units were delineated to represent relatively homogeneous, hydrologically continuous waterbodies. As such, multiple monitoring stations are expected to be representative of the water quality as part of the same "block," "sampling unit," or "treatment" as much as can be controlled under the requirements to use all available data provided to DEQ for the 303(d) assessment.	
1.8	5	Since each grab sample may largely reflect instantaneous ambient conditions in running water, I am also not sure if the grab samples with a large sampling time interval adequately reflect aquatic life criteria.	Instantaneous grab samples do not adequately resolve cyclical trends or high variability in water quality parameters. This leads to difficulty in assessing attainment of water quality criteria expressed as multi-day averages using grab sample data. The allowance of a non-zero exceedance frequency of 5%-10% is one way that the uncertainty in the through-time representativeness of grab samples is accounted for in the assessment process.	None required.
1.9	2,3	If the binomial approach is ultimately adopted for use by DEQ, DEQ must continue to do exploratory data analysis as part of implementing a binomial approach. DEQ does not describe how additional data analyses may be done alongside of, or as a precursor to, use of	DEQ applies reasonable quality assurance and quality control measures to the raw data before it is included for the assessment. DEQ is required to consider all readily available data that meets reasonable quality assurance requirements from other entities,	

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
		the binomial approach. Understanding the full nature of the raw data set is especially important in order to identify unusual patterns in the data, the quality of the data, and the need for additional monitoring to clarify the true condition of a waterbody.	including government agencies and the public, and is therefore limited in its ability to conduct additional monitoring prior to analyzing the data set. This data is used to assess attainment of waterbodies with the water quality criteria.	
1.10	2	The binomial distribution relies only on the proportion of exceedances, not their magnitude. Measurements that only slightly exceed a criteria are not distinguished from those that exceed a criteria by a large amount. There is much to be learned about waterbody conditions and variability in data sets, including sources of variation and potential outliers/spurious results, by examining the actual concentration measurements.	The 303(d) assessment process documents exceedances of water quality standards as thresholds. This process is concerned mainly with identifying waterbodies where pollutant concentrations exceed the thresholds. Once this has been determined to occur in a waterbody, the TMDL process provides more detailed analysis and modeling of the variability and magnitude of those exceedances, potential sources, and magnitudes of pollutant concentrations in preparing load allocations.	None required.
1.11	3	A hierarchical model, such as a generalized linear mixed effect model (GLMM) would allow DEQ to use unbalanced data sets in a way that would be much less dependent on the sampling design. The parameter of interest would be the assessment unit mean exceedance rate, rather than the exceedance rate weighted by sample size from each monitoring location. A model of this type is a natural extension of a one sample	DEQ is not aware of any states currently applying a generalized linear model to 303(d) assessment, and the method does not appear to have been reviewed in any EPA guidance. Just as the binomial method represented an improvement over earlier absolute threshold and raw score assessment methods that is now gaining wider adoption, the GLMM may be a refinement to the statistical assessment	Refer to staff for future consideration.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
		binomial model, but incorporates additional levels to deal with multiple comparisons (assessment units) and unbalanced sample designs within assessment units.	methodology that DEQ may consider in the future.	
1.12	3,5	A Bayesian approach provides a mechanism to incorporate the expert judgement and knowledge as a prior distribution and the assessment will be based on both the prior knowledge and the available data.	While adding Bayesian inference to the binomial method was proposed in the earliest introductions of the method (Smith et al, 2001), DEQ is not aware of any states, nor any EPA guidance, for its adoption. Selecting Bayesian prior probabilities is subjective and would leave DEQ open to challenge. DEQ has been criticized by stakeholders for relying too much on subjective expert judgement and desires to adopt a more fully data-driven and transparent methodology for 303(d) assessment.	

Question 2

2. Please comment on the selection of a confidence level of 90% for application of the binomial test for assessment purposes.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
2.1	1,2,4,5	The 90% confidence level is sufficient for regulatory purposes and falls within range of other states' application of the binomial method for water quality assessment.	Comment acknowledged.	None required.
2.2	6	Inherent variability in aquatic systems should be acknowledged in an 80% confidence level as a realistic and reasonable goal. An 80% confidence interval will provide a secondary benefit of results that are more consistent as one moves between the >1-in-3-year small sample size approach and the binomial approach of larger samples sizes.	California and Texas apply a confidence level of 80% to assessment with the binomial test. Texas varies the confidence level according to sample size. Nine other states that apply the binomial test use a confidence level of 90% Adoption of an 80% confidence level would result in listing with 1 excursion for sample sizes up to 30, instead of the current 18- this would significantly increase the number of waterbodies that are considered impaired when only one excursion is detected and sample sizes are low.	Refer to staff and stakeholder workgroup. Additional comparison of 80% and 90% confidence levels on error probabilities in the listing whitepaper.
2.3	2,4	DEQ could do more to explain its selection of 90% confidence. A typical range of 80%-95% percent is used by other states. An alpha value of 0.05 (95% confidence) is considered the standard for scientific research. For regulatory purposes 0.10 (90% confidence) is sufficient. With low sample size, an alpha level of	DEQ selected the 90% confidence level as a compromise between ensuring higher certainty in placing waters that are impaired on the 303(d) list and making assessing determinations with small data sets. Setting confidence levels too high (i.e. >95%) may actually increase Type II error rates by reducing the likelihood of listing impaired waters. While setting confidence too low (i.e. less than 80%) would result in additional listings that would not otherwise be	Refer to staff. Further evaluation added to listing methodology whitepaper.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
		0.05 could be too strict.	warranted, or removing waters from the 303(d) list when they should be considered impaired.	
2.4	2	DEQ should consider conducting an evaluation of other confidence level options to help ensure the long term needs of their water quality program are optimized.	There is no objective method for selecting an ideal confidence interval. Selection of a confidence interval is done *a priori* (McBride, 2005) and will directly affect the number of listings that will result. Selecting a lower confidence limit will increase the frequency of listings and type-I errors, and selection of a higher confidence interval will decrease the frequency of listings and increase type-II errors, relative to analysis of data from the same data set.	Refer to staff. Additional evaluation of the effect of selection of confidence levels on error probabilities in the listing whitepaper.
2.5	2,4	Ultimately, the choice of confidence level is an environmental policy/risk management decision that should reflect Oregon's level of risk adversity that can only be partially informed by science and technical information.	Choosing a confidence interval range is an inherently subjective process, but there is a range of commonly acceptable values used for hypothesis testing (McBride, 2005). DEQ has determined that a 90% confidence level is expected to balance program needs for accuracy while remaining consistent with the range of EPA guidance and best practice. Using 85% or 90% confidence levels are the most defensible values based on EPA guidance and state-wide best practices.	Refer to staff and advisory committee.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
2.6	5	For the purpose of protecting water resources, it will be costlier to have a high type II error rate in the watershed water quality assessment. Using a 90% confidence level (potentially larger type I error) may help to increase statistical power and reduce type II error. However, it is not clear if it is more effective than increasing sample size since both the complexity of the watersheds and effect size for pollutants may not be well quantified for many watersheds.	DEQ seeks to avoid both unnecessary economic and opportunity costs incurred to the regulated community and DEQ's TMDL program, by overestimating the number of impaired waters through type-I errors, and unnecessary costs to the environment and beneficial use of waters incurred by failing to identify impaired waters through type-II errors. The binomial listing methodology should encourage collection of larger data sets that will increase certainty in making impairment decisions for the 303(d) list and reduce errors relative to the current methodology.	None required.
2.7	3	Because DEQ estimates there will be >8000 assessment units, if even 5% of assessment units have data for some parameters, DEQ will probably perform several thousand binomial tests for each Integrated Report. The expected number of false positives (type I error rate * number of tests) may be high relative to the number of 'discoveries' (i.e., the number of rejected null hypotheses for all comparisons performed). For example, if DEQ performs 1600 (8000 assessment units * 5% with data * 4 parameters each on average) binomial tests at α = 0.1,	DEQ's proposed application of the binomial method increases the certainty required to place waters on the 303(d) list relative to the status quo. Most impaired waters have a proportion of excursions far above the nominal critical number of samples defined by the critical exceedance rate. These are less likely to be false positives than samples with only the nominal number of excursions required to list. The highest probability for error lies in small samples where the number of excursions places the proportion of expected exceedances within the confidence interval for the sample size. To date, DEQ's Category 5 listing rate is	None required. Refer to staff for future consideration.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
		the expected number of false positives is 160 assuming each test is independent and all exceedance rates are equal to the acceptable regulatory critical exceedance rate. The ratio of the expected number of false positives to the overall number of rejected null hypotheses is called the false discovery rate (FDR). For example, if DEQ rejected 200 null hypotheses when performing this many tests, then the FDR would be 160/200 = 80%. DEQ may be spending most of its time writing TMDLs for type-I errors.	approximately 18%, or 3495 of 19,421 segment x parameter combinations assessed. If the number of false positives were restricted to 5%, the false discovery rate would be estimated as 971/3495, or ~28%. The method used to assess those waters is strongly biased toward Category 5 results if any samples exceed the criteria, and has no quantifiable false positive rate. Therefore, the false discovery rate could either increase or decrease, but the overall number of listings would be expected to decrease with adoption of the binomial assessment method. Even though listing errors are undesirable, adopting the binomial method would likely reduce the number of false-positive listings referred to the TMDL program. Although not a replacement for accurate assessment of impaired waters, the TMDL process provides an additional opportunity for analysis that can confirm impairments or identify errors.	
2.8	3	Current statistical practice when performing many tests or comparisons, usually involves some consideration of the overall error rates when performing so many tests. For strict control of a single error, there are procedures that control the family wise error rate (FWER). For DEQ’s purposes, these procedures are too strict. However, there are	DEQ is interested in methods to refine the accuracy of water quality assessments. However, DEQ does not have an effective way to estimate the likely proportion of impaired waters independent of an assessment methodology such as the binomial. DEQ is not aware of any states currently applying an error correction to assessments and there is no standing guidance from EPA. While reducing the number of waters that are	Add reference in listing white paper and refer to staff for further discussion.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
		procedures that can control the FDR discussed above. The Benjamini-Hochberg procedure is simple to calculate using the observed p values and allows an analyst to control the false discovery rate for all the tests performed. This would allow DEQ to control the total proportion of assessment units that are potentially declared impaired erroneously, rather than only controlling the type I error rate in a single analysis. In either case, I strongly recommend DEQ consider the effect of its choice of type I error rates on the overall assessment methodology. This requires an assessment of the number of tests performed, the individual error rates, likely proportions of impaired versus unimpaired water bodies, and the appropriate error rate to control (e.g., FDR, FWER, etc). Because DEQ will be performing so many tests, ignoring the multiple comparison aspect may mean that the overall listing method will perform terribly despite the reasonableness of the approach to evaluating one	incorrectly identified as impaired is a goal, reducing the number of type-I errors will also lead to an increase in the type-II errors. In the case of 303(d) assessment there is an environmental cost that is incurred if type-II errors increase. Namely, DEQ would fail to identify waters that are actually impaired. McDonald (2014) suggested that if there is a cost to increased type-II errors, researchers may not want to correct for false positives.	

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
		parameter within one assessment unit.		

Question 3.

Please comment on the validity of the proposed null hypotheses and critical exceedance rates:

Where r = the true proportion of sample excursions in the waterbody, and
P1 = the acceptable regulatory critical exceedance rate, and
p2 = the unacceptable regulatory exceedance rate, for a desired effect size of 15%.

	Toxic Pollutants (chronic criteria)	Conventional Pollutants
Listing	H_O: $r \leq p_1 = 0.05$, H_A: $r > p_2 = 0.20$	H_O: $r \leq p_1 = 0.10$, H_A: $r > p_2 = 0.25$
Delisting	H_O: $r > p_1 = 0.05$ H_A: $r \leq \mathrm{p}_2 = 0.20$	H_O: $r > p_1 = 0.10$ H_A: $r \leq \mathrm{p}_2 = 0.25$

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
3.1	1,2,6	DEQ's proposed null hypotheses and critical exceedance rates seem valid and in line with EPA recommendations and practices in other states.	Comment acknowledged.	None required.
3.2	2,4	DEQ should further explain its justification for the 10% and 5% critical exceedance rates. Oregon could estimate sources and magnitude of variability in assessment data to help support their selection of critical exceedance rates.	The selection of critical exceedance rates are mainly dictated by EPA guidance as determined to reflect the duration and frequency component of water quality criteria that are established in water quality standards. As such, DEQ anticipates there is limited ability to change the allowable	Expanded justification for selection of the 5% and 10% exceedance rates in the whitepaper.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
			proportion of excursions unless they are recommended to be more stringent. A 5% rate for toxic substances and a 10% rate for conventional pollutants is almost universally applied by other states.	
3.3	6	Suggest slightly lower alternate hypothesis thresholds of 0.15 and 0.20 frequencies.	Comment acknowledged.	Refer to staff for evaluation.
3.4	2	The effect size of 15% seems to be supported by EPA guidance and use in other states.	Comment acknowledged.	None required.
3.5	2	The approach of using different critical exceedance rates in the same hypothesis test for Ho and Ha also seems useful.	Comment acknowledged. DEQ would like to point out disagreement among panelists in the following comment. Please see the next comment, below, for more explanation.	None required.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
3.6	3,4	DEQ is using a non-standard presentation of the null and alternative hypothesis and should revise. The alternative hypothesis must be framed in terms of p_1, the null value under the null hypothesis. If the null hypothesis for the test of a proportion is either H_O: $p \leq p1$ or $p = p_1$, then the alternative must be either H_A: $p \neq p_1$ (i.e., a two sided test) or H_A: $p > p_1$ (i.e., a one sided test).	EPA (2002) followed the suggestion in Smith et al. (2001) to apply a procedure to balance error rates at the desired effect size (i.e. 0.15). In this procedure the investigator specifies both an α and β level, a priori. Because of the discrete nature of the binomial distribution, α and β values can only be specified by identifying a minimum number of exceedances required to reject H_O for a given sample size. DEQ followed the example in Smith et al. (2001) and EPA (2002) for conventional pollutants to set an H_O: p_1=0.10. Smith et al. proposed that a population exceedance rate of 0.25 would indicate the rate of exceedances an agency would almost always want to ensure it was able to detect, and recommended specifying Ha: p_2=0.25. If this were used in listing decisions, waters with less than 10 percent exceedance would not be listed while waters with exceedance frequency above 25 percent would always be placed on the section 303(d) list. Waters that fall between these two values would sometimes be listed. This is equivalent to specifying a minimum effect size of 15%. A detailed explanation of the application of this procedure to determine critical values for the number	Updates to the listing methodology white paper.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
			of excursions for listing can be found in SWRCB (2005).	
3.7	1	It was not clear how the 15% effect size would be used. Within this 15% effects size between the regulatory rate and listing threshold (weak statistical power) it seems to fall short of the listing criteria but exceeds the regulatory exceedance criteria.	EPA recommends using an effect size of 15% to determine if there is a significant difference in the proportion of samples above the regulatory critical exceedance rate indicating impairment. In a test for proportions, the difference between the observed proportion and the regulatory proportion (the critical exceedance rate) is an effect size. In DEQ's proposal, for toxic substances the regulatory, or attaining threshold, is 5% and the listing threshold, is 20%, reflecting the desired effect size of 15%. Some states choose to simultaneously evaluate whether the proportion of excursions in a set of waterbody samples indicates the waterbody is above both the regulatory threshold and the listing threshold. DEQ has not proposed to apply this method at this time. The type-II (listing) error probabilities (1-β) graphed in attachment 2 show that for sample sizes above 50, there is sufficient power to reduce type-II errors to less than 5% for sample sizes >50 if an effect size of 15%	Clarification in listing methodology white paper

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
			is assumed. While the calculation of β is based on the desired effect size, it does not affect the selection of critical values used to determine impairment.	
3.8	3,4	DEQ should make clear that the "effect size" here is only being used to investigate the statistical error rates for a proposed statistical analysis.	In calculation of the critical values for a range of sample sizes, DEQ used only the regulatory exceedance rate (p_1). The listing exceedance rate (p_2) was used to calculate the type-II error probabilities (β) that would be expected.	Clarification in listing methodology white paper and attachment 1 procedure.

Question 4A

4a. In your professional opinion, are the selected hypothesis tests and Type I error rates sufficient to minimize environmental risk from making errors in conclusions to list or delist waterbodies?

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
4a.1	**1,2,3,4,5**	The expected type I error rates are suitable. The hypothesis tests and expected type I error rates are an improvement in listing and delisting decisions.	Comments acknowledged.	None required.
4a.2	**3**	DEQ has made great progress in proposing a listing method that is much improved from the current method. This test level is less conservative than the often used 5% error rate, which I think is appropriate given the high negative consequences of water pollution.	Comment refers to the alpha level of 0.10, or 90% confidence, proposed by DEQ. Please also see responses 2.1 – 2.3 above.	None required.

Question 4B

4b. In your professional opinion, does the type II error rate for the selected critical values for listing and delisting require balancing?

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
4b.1	**4**	The Type II errors appear reasonable.	Comment acknowledged.	None Required.
4b.2	**2,3,6**	Selecting the tolerable levels of both Type I and Type II decision errors is a choice to be made by the risk manager. I do not think that Type I and Type II error rates require balancing.	The EPA Consolidated Assessment Guidance (EPA, 2002) applied the error balancing approach proposed in Smith et al., 2001, to the section 303(d) listing process. EPA noted that balanced decision error rates are less affected when switching the assumption of the null hypothesis for listing and delisting, leading to more consistent decisions. Only the State of California appears to have fully implemented the error balancing approach in its application of the binomial test. DEQ did not include an error balancing approach in this proposal but set up a tool for calculating the critical value that could incorporate error balancing if recommended.	None required.

4b.3	**2**	Type II errors can be reduced by a) lowering the threshold of evidence required to list a waterbody, and b) basing the estimate of Type II errors on a higher critical exceedance value derived from the ability to distinguish between waters that are truly attaining versus waters that are truly impaired, i.e., the effect size.	DEQ's proposal to use an alternate critical exceedance rate for p2, the impairment threshold, would address b). Reducing the confidence interval from 90% to 80% would implement a). Please also see our response in 2.2 – 2.4 above.	Revisions to binomial white paper.
4b.4	5	For the purpose of protecting water resources, it will be costlier to have a high type II error rate. Using a 90% confidence level (potentially larger type I error) may help to increase statistical power and reduce type II error. However, it is not clear if it is more effective than increasing sample size since both the complexity of the watersheds and effect size for pollutants may not be well quantified for many watersheds.	DEQ recognizes the increased environmental risk of failing to identify impairments as reflected in type-II errors. By retaining the >1-sample-in-3-year critical exceedance rate for smaller sample sizes, DEQ counteracts the probability of making these types of errors less than 18, DEQ However, DEQ also seeks to be as accurate as possible seeks as much accuracy in listing waterbodies The type-II error rate is determined by the confidence level and sample size. Increasing sample size is the most immediate and effective way to reduce the type-II error.	Add analysis of type-II error rates versus sample size to the white paper.

Question 4 C

4c. If so, suggest alternative methods for calculating critical values while balancing the Type I and Type II error rates.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
4c.1	2	DEQ should consider more fully evaluating the consequences of their proposed and alternative error rates in order to affirm or alter their current choices.	The Type-I error rate is set by selection of the confidence level of the tests. The range of defensible error rates is 20%-5%. The type-II error rate is mainly determined by the confidence level and sample size. Type II errors will be greatly reduced with larger sample sizes. Sound statistical practice is to select the error rates *a priori* based on accepted scientific practice and policy consideration of tolerance for the risk associated with making each type of each error. It would introduce bias if DEQ were to select the parameters after evaluating the number of listings produced under each scenario and selecting the hypothesis test parameters based on optimization of some number of listing results.	Revisions and further analysis in binomial white paper.
4c.2	3	DEQ needs to be more consistent and clear in its discussion/presentation of Type I and Type II errors, especially as it relates to proposed use of two null hypotheses, one for listing and the reverse for delisting.	Type I and Type II error rates are relative to errors in rejecting the null hypothesis. Type I error does not always correspond to errors in listing. DEQ will revise the discussion about error rates to be more clear.	Revisions to whitepaper.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
4c.3	3,4	When there is too little data to make decisions with any level of confidence, the solution lies in prioritizing the collection of more information and using models that leverage all available information. For example, a Bayesian hierarchical model that models all assessment units within an area (e.g., ecoregion or watershed) for one parameter would allow DEQ to borrow power from nearby assessment units as suggested in Smith et al. 2001.	DEQ is recommending expanding the use of Category 3B and establishing protocols for follow up monitoring to improve correct identification of impairments where there is a high level of uncertainty in the data. DEQ is also considering using a multiple lines of evidence and overwhelming evidence approach for Category 5 listings, which would provide additional information about impairment status beyond the numeric evidence. This process is separate from the numeric listing methodology, but would provide DEQ a way to leverage additional information without the more complicated calculations of formal Bayesian analysis.	See discussion on category 3B and overwhelming evidence in the updated assessment methodology document.
4c.4	3,4	Collecting additional data incurs a trivial cost compared to the economic cost of preparing and implementing a TMDL to the state and regulated entities. It is the only way to reliably differentiate assessment units that are actually impaired from those that are only determined to be impaired because of statistical uncertainty.	DEQ agrees that better decision-making results from use of more complete data on waterbody condition. However, additional sampling or re-sampling to strengthen statistical reliability of the assessment is not always possible. This is partly due to resource constraints on DEQ's monitoring capacity, and partly due to the requirement to review all readily available data from public sources, or data submitted by the public. DEQ partially intends the adoption of	None required.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
			the binomial test to incentivize 3rd party entities to collect and submit more monitoring data; because it reduces the bias toward making false listing errors as sample sizes increase that is inherent in the current methodology.	

Additional Comments

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
A.1	2,6	Various minor or specific edits to the Listing or Delisting whitepapers.	Thank you for your comments. Edits provided will be incorporated in the next draft of the subject white papers.	Revised text in the whitepapers.
A.2	4	I checked the spreadsheet calculations for conventional listing for n equal 2-38 and found them to be consistent with my calculations.	Thank you for confirming the accuracy of the calculations in the spreadsheet. DEQ will follow the same procedure described in the spreadsheet to calculate the critical values to be used for listing and delisting purposes. This will reflect any changes due to final selection of the confidence interval or critical exceedance rates indicated by the peer review panel.	None required.
A.3	4	Check the figures in the Appendix 1 document. The conventional listing plot does not appear consistent with the figure in the spreadsheet.	The figures in the spreadsheet represent the final calculation of alpha and beta values for parameters given in the hypothesis test. If the plots do not match, the plot in Attachment 1 is in error.	Update figure in Appendix 1 with the figure in the spreadsheet.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
A.4	**2**	DEQ should consider data from outside the most recent 3-year assessment period to ensure that opportunities to make more accurate assessment decisions using larger data sets are not missed.	DEQ's new assessment units will potentially combine data from multiple data sources that represent conditions within the waterbody. This will allow for larger datasets than were reasonably expected to be collected at an individual monitoring station within a 3-year period. Data from the most recent 3 years is expected to better represent current conditions within waterbodies, and will allow DEQ to asses changes in water quality that occur since the last assessment was completed if new data has been collected. For the 2018 Assessment, DEQ will assess data from a 10-year window to encompass new data collected since the last assessment. For the 2020 assessment, DEQ may revisit consideration of the most suitable date range for assessment.	Referral to staff.
A.5	**2**	DEQ should consider potentially re-evaluating data supporting previous listings before adopting the assumption of impairment for all currently listed waterbodies.	DEQ anticipates to evaluate listings with any new or available data because of the 10-year data window being used for the 2018 assessment. This will in-effect be a re-evaluation of any listings made with data from 2007-2012 using the updated listing	Referral to staff. Adjustments to assessment methodology document.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
			methodology. However, older listings where a TMDL has not been completed or sufficient data has not been collected since 2007 will not be evaluated. DEQ is further considering new protocols for re-sampling and conformation of older listings that have not completed the TMDL process. However, as placement on the 303(d) list indicates waterbodies are legally considered to be impaired, even if listed under a different assessment method, the impaired status is assumed unless there is new and sufficient data available to show attainment.	
A.6	**2**	DEQ should improve and supplement the existing binomial calculator spreadsheets to improve the transparency and broad understanding of their intended uses of the binomial distribution:	DEQ intended the binomial spreadsheets as a demonstration of the procedure used for calculating the tables of critical values relative to sample size for the benefit of the review panel. We consider the tables to be a more transparent and effective tool to communicate the thresholds used for determining impairment of aquatic life using the numeric criteria to the public. The spreadsheet was used to illustrate the calculations used to derive these	Updates to white paper and assessment methodology.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
			tables of critical values, but is not intended as a calculator for use by the general public. Any changes to the specific test parameters in the spreadsheet would change the critical value thresholds for determining impairment.	

Cited References

EPA. (2002). *Consolidated Assessment and Listing Methodology Towards a Compendium of Best Practices First Edition*. U.S. Environmental Protection Agency, Office of Wetlands, Oceans, and Watersheds.

McBride, G. B. (2005). *Using statistical methods for water quality management: issues, problems, and solutions.* Hoboken, New Jersey: John Wiley and Sons, Inc.

McDonald, John H. (2014). *Handbook of Biological Statistics* (3rd ed.). Baltimore, Maryland: Sparky House Publishing.

Smith, E. P., Ye, K., Hughes, C., & Shabman, L. (2001). Statistical assessment of violations of water quality standards under Section 303(d) of the clean water act. *Environ Sci Technol*, *35*(3), 606–612.

SWRCB. (2005). *Final functional equivalent document. Water quality control policy for developing California's clean water act section 303(d) list.* California State Water Resources Control Board.

Section 3 Direct Peer Review Responses

SCIENTIFIC PEER REVIEW: SOLICITATION REQUEST FORM

Reviewer Information	
Reviewer Name:	**Title:**
Email Address:	**Contact Phone #:**
Employer:	**Employer Category:** (federal agency, state agency, academic, professional organization/consultant)
Subject Matter: Exact Binomial Assessment Methodology	

Purpose of Review & Specific Action Required: DEQ is soliciting independent scientific and technical input regarding the binomial test that is being proposed for Clean Water Act section 305(b) and 303(d) assessment purposes in the 2018 Integrated Report. DEQ is proposing to apply the exact binomial statistical test to chronic aquatic life toxics criteria and conventional pollutants (i.e. dissolved oxygen, pH etc.) for assessment purposes. DEQ is not proposing to apply the binomial for assessment of acute standards or human health criteria. Please provide review comments on the questions below.

1. **Is DEQ's proposed use of the exact binomial statistical test valid and defensible for assessment of:**
   - **a) chronic aquatic life toxics criteria?**
   - **b) conventional pollutant criteria?**

2. **Please comment on the selection of a confidence level of 90% for application of the binomial test for assessment purposes.**

3. **Please comment on the validity of the proposed null hypotheses and critical exceedance rates:**

   Where r = the true proportion of sample excursions in the waterbody, and

   P1 = the acceptable regulatory critical exceedance rate, and

   p2 = the unacceptable regulatory exceedance rate, for a desired effect size of 15%.

	Toxic Pollutants (chronic criteria)	**Conventional Pollutants**
Listing	H_O: $r \le p_1 = 0.05$, H_A: $r > p_2 = 0.20$	H_O: $r \le p_1 = 0.10$, H_A: $r > p_2 = 0.25$
Delisting	H_O: $r > p_1 = 0.05$ H_A: $r \le p_2 = 0.20$	H_O: $r > p_1 = 0.10$ H_A: $r \le p_2 = 0.25$

4. **Please comment on the methodology for calculation of critical values for listing and delisting, detailed in Attachment 1 and Attachment 2.**
   - a. **In your professional opinion, are the selected hypothesis tests and Type I error rates sufficient to minimize environmental risk from making errors in conclusions to list or delist waterbodies?**
   - b. **In your professional opinion, does the type II error rate for the selected critical values for listing and delisting require balancing?**
   - c. **If so, suggest alternative methods for calculating critical values while balancing the Type I and Type II error rates.**

Timeline for Review Completion: Reviews should be completed and returned electronically to DEQ by January 29, 2017.

DEQ Point-of-Contact for Reviewer	
DEQ Contact Name: Becky Anthony	**Title: Interim Integrated Report Coordinator, Oregon DEQ**
Email Address: anthony.becky@deq.state.or.us	**Contact Phone #: 541-686-7719**

Specific instructions for providing review comments to DEQ:

Reference documents attached to this request are: (1) Attachment 1- Binomial test procedures; (2) Attachment 2- Binomial critical value tables calculations; and (3) Listing and Delisting Methodology Whitepaper.

DEQ staff are available to answer questions, provide additional information or clarifications. Questions should be directed to Becky Anthony (see contact information above).

Please provide peer review comments to DEQ electronically to integratedreport@deq.state.or.us by January 29, 2017.

DEQ follow-up and use of review comments:

DEQ will compile all of the comments received and may reach out to reviewers for explanatory purposes. Comments will be summarized and used to inform revisions to Oregon's assessment methodology.

Comments on subject matter reviewed (please attach additional pages as needed)**:**

Comments (continued)