State of Oregon Department of Environmental Quality

Summary of Binomial Listing Methodology Peer Review

Memorandum

To: Integrated Report Work Group
Date: March. 14, 2018
From: Integrated Report Improvement Team
Subject: Summary of binomial listing methodology 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[1]. Selection of a confidence level within the accepted range of 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

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

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	None required.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
			would allow for more accurate characterization of the 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	Refer to staff and advisory committee for consideration.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
			exceedance rates to the assessment of acute toxics criteria as well as to chronic toxics criteria. 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	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.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
		simply because a well-established statistical method is used.		
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 regularly encountered when performing these binomial tests.	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 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.	Revisions to assessment unit white paper.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
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 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.	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, 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.	

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
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 binomial model, but incorporates additional levels to deal with multiple comparisons (assessment units) and unbalanced	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 methodology that DEQ may consider in the future.	Refer to staff for future consideration.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
		sample designs within assessment units.		
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	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
		low sample size, an alpha level of 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	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.	None required. Refer to staff for future consideration.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
		1600 (8000 assessment units * 5% with data * 4 parameters each on average) binomial tests at α = 0.1, 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.	To date, DEQ's Category 5 listing rate is 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.	

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
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 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	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 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.	Add reference in listing white paper and refer to staff for further discussion.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
		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 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	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.	Expanded justification for selection of the 5% and 10% exceedance

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
		their selection of critical exceedance rates.	As such, DEQ anticipates there is limited ability to change the allowable 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.	rates in the whitepaper.
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	Updates to the listing methodology white paper.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
			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 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.	Clarification in listing methodology white paper

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
			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% 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	Revisions and further analysis in binomial white paper.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
			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.	
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.
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	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	See discussion on category 3B and overwhelming evidence in the updated assessment

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
		or watershed) for one parameter would allow DEQ to borrow power from nearby assessment units as suggested in Smith et al. 2001.	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.	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 the binomial test to incentivize 3rd party entities to collect and submit more monitoring data; because it reduces the bias toward	None required.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
			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.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
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.
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	Referral to staff.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
			consideration of the most suitable date range for assessment.	
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 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	Referral to staff. Adjustments to assessment methodology document.

COMMENT	PANELIST	COMMENT SUMMARY	RESPONSE	REVISION
			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. We consider the tables to be a more transparent and effective tool to communicate to the public the thresholds used for determining impairment of aquatic life using the numeric criteria. The spreadsheet was used to illustrate the calculations used to derive these tables of critical values, but is not intended as a calculator for use by the general public. Any changes to the specific test parameters would change the critical value thresholds for determining impairment.	Updates to white paper and assessment methodology.

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.