State of Oregon Department of Environmental Quality

Peer Review Materials

Date of Request: December 21, 2017

SCIENTIFIC PEER REVIEW: SOLICITATION REQUEST FORM

Reviewer Information	
Reviewer Name: Dr. Douglas McLaughlin	**Title:** Principal Research Scientist
Email Address: dmclaughlin@NCASI.org	**Contact Phone #:** 269-350-6158
Employer: National Council for Air and Stream Improvements	**Employer Category:** Professional organization (federal agency, state agency, academic, professional organization/consultant)
Subject Matter: Exact Binomial Assessment Methodology **Timeline for Review Completion:** Reviews should be completed and returned electronically to DEQ by January 29, 2018.	

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?

If the exact binomial test is not appropriate, what alternative method may be appropriate given the limitations described in the attachment?

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

If a 90% confidence level is not appropriate, what alternative confidence level may be appropriate given the limitations described in the attachment?

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 Pollutant (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: $r \leq p_2 = 0.20$	H_O: $r > p_1 = 0.10$, H_A: $r \leq p_2 = 0.25$

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.

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 December 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)**:**

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?

Question 1 (a and b addressed together)

Yes, overall ODEQ's use of the exact binomial method appears defensible, and represents an improvement in the basis for and transparency of, listing and delisting decisions relative to current practice. DEQ could further describe its justification for selecting 10% and 5% critical exceedance rates. Also, DEQ should reconsider use of the binomial approach for assessing attainment/non-attainment for acute WQC, an option that appears to be within EPA guidelines.

To ensure appropriate application of the binomial approach, DEQ must also keep in mind important assumptions associated with its use when developing monitoring programs or selecting data for assessment purposes. This includes 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.

DEQ must also combine its use of the binomial approach with more complete evaluation of the data being assessed so that more subtle patterns over time and space, and the potential for outliers, can be identified. Because the binomial test only evaluates the presence or absence of sample exceedances, it's appropriate use will be better informed by more thorough data analysis.

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

If a 90% confidence level is not appropriate, what alternative confidence level may be appropriate given the limitations described in the attachment?

Using a 90% confidence level is reasonable and falls within other applications of the binomial method for water quality assessment. However, the selection of a confidence level is not a purely scientific endeavor. It reflects the tolerance for decision errors which, as EPA's data quality objectives guidance points out, are ultimately in the hands of risk managers as they weigh multiple factors. To that end, DEQ could do more to explain its selection of 90%. As supporting documents point out, other values have been used elsewhere. DEQ should consider conducting and presenting a more detailed evaluation of other confidence level options to help ensure that the long term needs of their water quality program are optimized.

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 Pollutant (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: $r \leq p_2 = 0.20$**	**H_O: $r > p_1 = 0.10$,** **H_A: $r \leq p_2 = 0.25$**

DEQ's proposed null hypotheses and critical exceedance rates seem valid and in line with EPA recommendations and practices in other states. DEQ could more fully describe its rationale for the values selected, however. It appears that the Florida Department of Environmental Protection, for example, attempted to estimate sources and magnitude of variability in assessment data to help support their selection of critical exceedance rates. It is not clear that Oregon has the same type of analysis. Such an evaluation could show, for example, that the variability in toxicant measurements (e.g., as shown in split or duplicate samples or from other aspects of data quality assessment) is the same or higher than for conventional pollutants, leading to critical exceedance rates that are different than DEQ is currently proposing to use. The effect size of 15% seems to be supported by EPA guidance and use in other states. The approach of using different critical exceedance rates in the same hypothesis test for Ho and Ha also seems useful.

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?

In general, the methods used to calculate critical values for listing and delisting are reasonable, notwithstanding my related comments in Questions 1-3 and other more specific comments outlined in the additional pages that I am submitting. Also, the hypothesis tests and Type I error rates should provide for improved accuracy and transparency in listing and delisting decisions. As DEQ's analysis clearly shows, an important way to improve listing and delisting decision accuracy, and best serve DEQ's diverse water quality management goals, is through the use of larger, high quality data sets. Therefore, DEQ should not artificially limit the applicable data to a 3-year assessment window when

additional relevant data are available from previous assessment periods, and should continue to seek ways to collect data that are more spatially and temporally rich.

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

Type II errors can be thought of as “errors of inaction” with respect to the null hypothesis, i.e., not rejecting the null hypothesis when it should be rejected (leading to “false negative” errors). This applies to both listing and delisting scenarios in DEQ’s proposal. So for listing decisions, a Type II error occurs when a waterbody is not listed when it should be. 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. As EPA’s CALM document points out, selecting the tolerable levels of both Type I and Type II decision errors is a choice to be made by the risk manager who may be charged with taking many factors into consideration. Type I and Type II errors do not generally “require” balancing. DEQ should consider more fully evaluating the consequences of their proposed and alternative error rates in order to affirm or alter their current choices.

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

State of Oregon Department of Environmental Quality

Peer Review Materials

Date of Request: December 21, 2017

SCIENTIFIC PEER REVIEW: SOLICITATION REQUEST FORM

Reviewer Information	
Reviewer Name: Dr. Gerrad Jones	**Title:** Assistant Professor
Email Address: Gerrad.Jones@oregonstate.edu	**Contact Phone #:** 541-207-4534
Employer: Oregon State University	**Employer Category:** Academic (federal agency, state agency, academic, professional organization/consultant)
Subject Matter: Exact Binomial Assessment Methodology **Timeline for Review Completion:** Reviews should be completed and returned electronically to DEQ by January 29, 2018.	

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?

If the exact binomial test is not appropriate, what alternative method may be appropriate given the limitations described in the attachment?

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

If a 90% confidence level is not appropriate, what alternative confidence level may be appropriate given the limitations described in the attachment?

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 Pollutant (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: $r \leq p_2 = 0.20$	H_O: $r > p_1 = 0.10$, H_A: $r \leq p_2 = 0.25$

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.

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 December 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)**:**

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?

For both the toxic and conventional pollutants, I think the binomial test is scientifically defensible. I think statistical tests remove some of the arbitrary nature of determining when a water body should be listed on the 303d list. I think the binomial test is particularly useful because it is a simple yes/no test, and there is little ambiguity as to whether or not a value is above or below a threshold concentration. When appropriate data assumptions are met, parametric statistical test always have more power than non-parametric tests. For example, the one sample t-test would be a parametric equivalent to the non-parametric binomial test. Given a variety of reasons (e.g., outliers, different sampling, handling, and analytical procedures, etc), I suspect a parametric test would be inappropriate. While statistics are useful, they can also be confusing even to those who have a working knowledge of statistical analyses. Therefore, I really appreciate the fact that this tool is so simple and can be put into a simple tabular format. This is a very big strength in my opinion.

Within the scientific community, an alpha value of 0.05 is considered the standard when evaluating statistical significance, although 0.10 is not uncommon. If I were to review a top tier journal article, I would accept nothing less than an alpha level of 0.05. For regulatory purposes, I think 0.10 is perfectly sufficient especially considering that small sample sizes are to be expected frequently. With low sample size, an alpha level of 0.05 could be too strict of a criteria. Related to the alpha, I think the exceedance rates are appropriate, but I don't know why the values are different for toxic and conventional pollutants. With 0.05, fewer exceedances are acceptable, thereby making it a more conservative approach. I suppose for toxics, a more conservative stance makes sense in order to protect ecosystem/human health, but why not make both 0.05? Nevertheless, I still think the binomial test for this purpose is scientifically defensible as is.

Sample size is always a problem, and for the lowest number of exceedances (2), there is little statistical power here. I think this "bin" was for a sample size of up to 18 samples. It is noted that a large proportion of the datasets will fall into this bin, which is unfortunate given the low statistical power in this range. This is the reality of the situation, and I think the description provided acknowledges this and I think the usage of the binomial test in this range is well justified by DEQ. Despite this limitation, I still think the binomial test is appropriate.

Finally, It was not clear how the 15% effects size would be used. It seems to be merely supporting information to the test, which is not bad. It seems that by including this information into the alternative hypothesis, there is a 3rd option: 1) below the regulatory rate (strong statistical power: do not reject null hypothesis), 2) above the listing threshold (strong statistical power: reject null hypothesis), between regulatory rate and listing threshold (weak statistical power). In this last scenario, it is within this 15% effects size, which seems to fall short of the listing criteria but exceeds the regulatory exceedance criteria. I acknowledge that I do not have all the information about how this will be used, but it is something that I was not able to resolve on my own and still remains an open question.

Overall, I think the binomial test as presented herein is a scientifically defensible test for determining whether or not a water body should be listed/delisted from the 303d list.

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

If a 90% confidence level is not appropriate, what alternative confidence level may be appropriate given the limitations described in the attachment?

Please response to question 1. In summary, I think the 90% confidence level is sufficient for regulatory purposes and is the minimum level appropriate for scientific/research 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 Pollutant (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: $r \leq p_2 = 0.20$**	**H_O: $r > p_1 = 0.10$,** **H_A: $r \leq p_2 = 0.25$**

lease response to question 1. I think these are appropriate. The toxics threshold exceedance rate is lower than the conventional pollutants. This makes sense if you are trying to take a more strict stance on toxics, but why not be strict on both?

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?

In my opinion, I believe this is a good place to start, and transitioning to this statistical procedure is logical and well justified. I think the results from the binomial test will help interpret the data: it is simple and straightforward. It is difficult to know if the error rates and the selected alpha value will ultimately achieve the desired water quality goals, but I think they are suitable for minimizing errors within the decision making process.

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

I am unclear how the type II error rate will be used in this analysis. Based on the excel file, the allowable number of samples that can exceed the regulatory limit is entirely dependent on the chosen alpha level. Looking through the excel table, I saw no evidence to indicate how the type II error rate had any bearing on the data that was presented in the tabulated values. Therefore, I am not certain how the type II error rate or the effect size affects listing and delisting. This could be an oversight on my part, but because I don't know how this value affects the calculation for the tabulated data, I cannot make a suggestion for part c.

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

CITY OF PORTLAND
ENVIRONMENTAL SERVICES

Water Pollution Control Laboratory

6543 N Burlington Avenue, Bldg 217, Portland, Oregon 97203 ▪ Nick Fish, Commissioner ▪ Michael Jordan, Director

Technical Memorandum

TO: BECKY ANTHONY

FROM: JASON LAW

COPIES: PETER ABRAMS, LINDA SCHEFFLER

DATE: 1/29/2018

SUBJECT: SCIENTIFIC PEER REVIEW OF EXACT BINOMIAL ASSESSMENT METHODOLOGY

Summary

The Oregon Department of Environmental Quality (DEQ) is proposing to an exact binomial test to evaluate whether water quality samples from assessment units throughout the state to evaluate whether water bodies within those assessment units are exceeding chronic aquatic life toxics criteria and conventional pollutant criteria. This method is a vast improvement over the current method, however there are still two significant issues with the method. First, the method does not address the unbalanced data sets that will be regularly encountered when performing these binomial tests. The data "pooling" that DEQ is considering performing violates the independence and identically distributed assumption of the exact binomial test. Second, although DEQ has carefully considered error rates for individual tests, DEQ has made no mention of the many tests that will likely be performed and the effect of error rates for individual tests on the entire assessment procedure. This is not compatible with current statistical practice and is easily rectified by considering overall error rates for the listing methodology and controlling them. Finally, it is very difficult to evaluate a proposed data analysis without a close examination of the data to be analyzed or a discussion of the entire procedure (i.e., the many tests to be performed). I highly recommend that DEQ evaluate their proposed method using actual water quality data or simulated data that is generated using realistic assumptions informed by actual data (e.g., realistic estimates of variance components across time and space, numbers of assessment units, sample sizes of locations and numbers of samples at each location, likely proportion of impaired assessment units, etc). The issues I bring up can only be evaluated satisfactorily when the basic conditions under which the assessment methodology will be performed are identified and the method is evaluated considering these conditions.

Questions

1. *Is DEQ's proposed use of the exact binomial statistical test valid and defensible for assessment of:*
    a. *Chronic aquatic life criteria?*

The general approach of using a binomial model for assessing whether a water body's rate of exceeding a chronic aquatic life criteria is higher than acceptable is valid and defensible. Basing compliance decisions on inference about a proportion of exceedances, rather than a fixed >1-in-3-year exceedance irrespective of sample size, correctly makes the compliance decision about the unknown population parameter (i.e., the unobserved rate of exceedances for the entire water body over the applicable time period).

    b. *Conventional pollutant criteria?*

Similarly, using a binomial model to assess decisions about compliance with a conventional pollutant criteria correctly makes the decision about the unknown parameter of interest – the unknown population exceedance rate.

However, in both the chronic aquatic life criteria and conventional pollutant criteria cases, it is impossible to evaluate the choice of the exact binomial test, without evaluating the sampling context. The exact binomial test, like other one sample, simple hypothesis tests assumes that the data is independent and identically distributed[1,2]. If DEQ adopts the proposed assessment units for 303d listing purposes, this assumption is unlikely to hold. Datasets used for the Integrated Report incorporate any data available. The available datasets are almost always unbalanced with respect to sampling locations and samples collected over time. DEQ's proposed method[3] to deal with unbalanced data sets would produce data sets that pool samples collected from multiple locations. Not only does this violate a basic assumption of the exact binomial test[4], it is trivial to create datasets from simple "synthetic" watersheds which meet the acceptable exceedance rate specified in the attachment, but which easily produce datasets that would be declared impaired using DEQ's proposed methodology. For example, an assessment unit that is comprised solely of the mainstem of a small river with two monitoring stations – one upstream and one downstream of a major point source. If we assume that the river upstream meets the relevant water quality standard while the river downstream exceeds the relevant water quality standard, then whether the assessment unit is declared impaired when applying the method proposed in Attachment 1 to the pooled data depends solely on the relative sample sizes of the two sites. Thus, the decision depends on the study design and not the state of the assessment unit! Simply increasing the sample size upstream of the point source would result in the assessment unit being declared unimpaired or delisted. Clearly, a more realistic "model"

---

[1] Smith, Eric P., et al. "Statistical assessment of violations of water quality standards under Section 303 (d) of the Clean Water Act." *Environmental Science & Technology* 35.3 (2001): 606-612.

[2] "Draft Consolidated Assessment and Listing Methodology: Toward a Compendium of Best Practices". Office of Wetlands, Oceans and Watersheds. April 20, 2001. U.S. EPA. 2001b.

[3] http://www.oregon.gov/deq/FilterDocs/WhitePaperIRdataAggreg.pd

[4] Smith, Eric P., et al. "Statistical assessment of violations of water quality standards under Section 303 (d) of the Clean Water Act." *Environmental science & technology* 35.3 (2001): 606-612.

would be that the average of the upstream of downstream conditions should be used as the parameter, which is not the same as the simple pooled upstream and downstream data.

A method must be used which allows for data sets that are unbalanced with regard to sampling locations and repeated samples over time. A hierarchical model, such as a generalized linear mixed effect model[5] (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.

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

DEQ has chosen a 90% one sided confidence level for assessing a single water quality parameter and its associated criteria (chronic aquatic life criteria or conventional pollutant criteria). To prepare an Integrated Report, DEQ must screen all relevant water quality parameters in each assessment unit for which it has sufficient data. 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 $\alpha = 0.1$, 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 hypothesis 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 1 errors.

Many of the assessment units declared impaired using this procedure, perhaps even the majority, could be in error using these nominal error rates. However, the actual rates in real data would depend on the actual exceedance rates encountered (e.g., if all the actual exceedance rates are > 10%, there would be no false positives and a false discovery rate of 0).

If a 90% confidence level is not appropriate, what alternative confidence level may be appropriate given the limitations described in the attachment.

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[6] 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

---

[5] Bolker, Benjamin M., et al. "Generalized linear mixed models: a practical guide for ecology and evolution." *Trends in ecology & evolution* 24.3 (2009): 127-135.

[6] Benjamini, Yoav, and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the royal statistical society. Series B (Methodological)* (1995): 289-300.

potentially declared impaired erroneously, rather than only controlling the type I error rate in a single analysis.

More complex hierarchical models like a Bayesian hierarchical model would allow for some of these tests to be collapsed into a single model, obviating the need to make power sapping corrections like the procedures mentioned above.

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 parameter within one assessment unit.

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%.*

The alternative hypothesis is stated incorrectly in each of the listing and delisting cases for toxic pollutants (chronic criteria) and conventional pollutants. If the null hypothesis for the test of a proportion is either H_O: $p \leq p_1$ or $p = p_1$, then the alternative must be either H_A: $p \neq p_1$ (i.e., a two sided test) or Ha: $p > p_1$ (i.e., a one sided test)[7]. The alternative hypothesis must be framed in terms of p_O, the null value under the null hypothesis.

An "effect size" is a term used to describe the magnitude of the difference between an observed effect and the null value: $\frac{\mu_1 - \mu_O}{\sigma}$ and can be estimated using the sample mean and sample standard deviation along with the null value. The effect size is useful to summarize the magnitude of the observed effect. A proposed effect size is also useful in study design and sample size calculation to calculate the power that a specific study design would have to detect an effect size of interest. For example, DEQ might calculate the power to detect a 20% exceedance rate to evaluate whether changes to a proposed monitoring design are necessary or to establish a minimum sample size for using a statistical hypothesis test to determine compliance with water quality standards.

However, the "effect size" is not used to set up the null and alternative hypotheses in a test for a proportion. In addition, power calculations are generally not used during a data analysis and "retrospective" or "post-hoc" power analysis is not acceptable[8]. DEQ should revise the null and alternative hypotheses and make clear that the "effect size" here is only being used to investigate the statistical error rates for a proposed statistical analysis.

4. *Please comment on the methodology for calculation of critical values for listing and delisting, detailed in Attachment 1 and Attachment 2.*

---

[7] Utts, Jessica M., and Robert F. Heckard. *Mind on statistics*. Cengage Learning, 2004.

[8] Hoenig, John M., and Dennis M. Heisey. "The abuse of power: the pervasive fallacy of power calculations for data analysis." *The American Statistician* 55.1 (2001): 19-24.

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?*

I think DEQ has made great progress in proposing a listing method that is much improved from the current method. I also think that a Type I error rate of 0.1 is an appropriate test level for minimizing environmental risk. 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.

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

I do not think that Type I and Type II error rates require "balancing". Unfortunately, DEQ is looking for a solution to the problem of too little data to have confidence in decisions in the wrong places. It is important to acknowledge that making both errors have environmental, public health and economic risks. "Balancing" both error rates does not solve any of these problems: by definition it just trades one for the other. And I would argue that environmental risk is not always decreased when the Type I error rate is increased. Increasing the Type I error rate to decrease the Type II error rate for a data set of a fixed size means that more assessment units will be declared impaired (both erroneously and correctly). However, this doesn't acknowledge the true state of nature. If impairment is a problem that is less common than not being impaired, then the number of erroneous listings could greatly outnumber the number of correct listings. For example, using the balanced error rates quoted in Smith 2001, if 10% of assessment units statewide are impaired and assuming exactly 4000 assessment units which are all assessed with a sample size of 10, then the expected number of type I errors would be 936 (i.e., 0.26 * 4000 * 0.9) at the error rate of 0.1, while the expected number of correctly classified impairments would be 304 ((1 – 0.24) * 4000 * 0.1). Over three times more assessment units would be declared impaired erroneously than correctly. Increasing the Type I error rate would still incur high environmental and economic costs: effort would be wasted on erroneous listings and true environmental problems would be lost amid many erroneous listings for which agencies have too few resources to address.

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[9]. 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 impaired from those that are only declared so when error rates are "balanced". And environmental risk can only be truly decreased when environmental problems are correctly identified and the limited resources available to address them are allocated to address these issues.

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

---

[9] Smith, Eric P., et al. "Statistical assessment of violations of water quality standards under Section 303 (d) of the Clean Water Act." *Environmental Science & Technology* 35.3 (2001): 606-612.

A hierarchical model such as a GLMM (fit via a Bayesian or frequentist approach) would allow DEQ to address unbalanced data sets within assessment units, incorporate additional information from nearby assessment units as proposed in Smith 2001, and to deal with the multiple testing issue pointed out in question 2[10,11]. 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 sample designs within assessment units. The "critical values" in a model like this would be posterior confidence intervals of the relevant model parameters. For example, a Bayesian 80% highest posterior density (HPD) interval that does not include 0.1 for a conventional pollutant would be equivalent to the exact binomial test approach proposed here for a one sample dataset.

---

[10] Gelman, Andrew, Jennifer Hill, and Masanao Yajima. "Why we (usually) don't have to worry about multiple comparisons." *Journal of Research on Educational Effectiveness* 5.2 (2012): 189-211.

[11] Gelman, Andrew, and Francis Tuerlinckx. "Type S error rates for classical and Bayesian single and multiple comparison procedures." *Computational Statistics* 15.3 (2000): 373-390.

State of Oregon Department of Environmental Quality

Peer Review Materials

Date of Request: December 21, 2017

SCIENTIFIC PEER REVIEW: SOLICITATION REQUEST FORM

Reviewer Information	
Reviewer Name: Dr. Jon Harcum	**Title:** Principal engineer/hydrologist
Email Address: jon.harcum@tetratech.org	**Contact Phone #:** 864-656-2541
Employer: Tetra Tech	**Employer Category:** Consultant (federal agency, state agency, academic, professional organization/consultant)
Subject Matter: Exact Binomial Assessment Methodology **Timeline for Review Completion:** Reviews should be completed and returned electronically to DEQ by January 29, 2018.	

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?

If the exact binomial test is not appropriate, what alternative method may be appropriate given the limitations described in the attachment?

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

If a 90% confidence level is not appropriate, what alternative confidence level may be appropriate given the limitations described in the attachment?

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 Pollutant (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: $r \leq p_2 = 0.20$	H_O: $r > p_1 = 0.10$, H_A: $r \leq p_2 = 0.25$

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.

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 December 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)**:**

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?

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. It is this reviewer's understanding that ODEQ is not proposing to use the methodology for acute criteria. To the extent that that monitoring data used in the evaluation are representative of the waterbody being sampled, ODEQ's proposed use of the binomial test is a valid and defensible methodology for evaluating chronic aquatic life toxics criteria and chronic conventional pollutant criteria. This position is further supported by other states (Anthony et al. 2017, Appendix 2, Table 9) that use similar methodololgies.

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

   If a 90% confidence level is not appropriate, what alternative confidence level may be appropriate given the limitations described in the attachment?

DEQ is proposing to use the 90% confidence level for assessment purposes. That is, ODEQ has computed the allowable number of exceedances so as to not allow the significance level to exceed 0.10 except for the smallest sample sizes. The value of 90% falls within the typical 80-95 percent range used by other states. Ultimately, the choice of confidence level is an environmental policy/risk management decision that should reflect Oregon's level of risk adversity. DEQ appears to have considered the pros/cons of different confidence levels in choosing the 90% confidence level.

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 Pollutant (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: $r \leq p_2 = 0.20$**	**H_O: $r > p_1 = 0.10$,** **H_A: $r \leq p_2 = 0.25$**

DEQ proposes to use p_1 to calculate α, and p_2 is used to calculate β. As provided in the reviewed information, only α, and consequently p_1, has an effect on the number of excursions to list or delist. Nevertheless, setting the desired effect size and computing β is a useful exercise for purposes of reviewing the overall assessment methodology. Nevertheless, I have two concerns. 1) The DEQ document does not present a sufficient rationale for setting p_1 to 0.05 and 0.10 for toxics and conventional pollutants, respectively. That is, why not 0.05 or 0.10 for both categories? 2) Although the approach adopted by DEQ for presenting statistical hypotheses has been used elsewhere, it is, nevertheless, not a standard strategy for presenting hypothesis statements. It is my recommendation that H_A, in all cases, be the simple reflection of H_O. For example, H_A for listing toxic pollutants would be H_A: $r > p_1$. Graphics, tables, or text presenting β, would be footnoted indicating the desired effect size was set to 0.15 greater than p_1.

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.

In my professional opinion, the selected hypothesis tests and Type I error rates represent a reasonable balance between environmental risk management, societal costs, and the limitations often associated with typical data sets. Further, the Type II errors appear reasonable.

Other comments.

1) Check the figures in the Appendix 1 document. The conventional listing plot does not appear consistent with the figure in the spreadsheet.

2) I checked the spreadsheet calculations for conventional listing for n equal 2-38 and found them to be consistent with my calculations.

State of Oregon Department of Environmental Quality

Peer Review Materials

Date of Request: December 21, 2017

SCIENTIFIC PEER REVIEW: SOLICITATION REQUEST FORM

Reviewer Information	
Reviewer Name: Patrick Moran	**Title:** Aquatic Toxicologist
Email Address: pwmoran@usgs.gov	**Contact Phone #:** 253-552-1646
Employer: USGS	**Employer Category:** Federal Agency (federal agency, state agency, academic, professional organization/consultant)
Subject Matter: Exact Binomial Assessment Methodology **Timeline for Review Completion:** Reviews should be completed and returned electronically to DEQ by January 29, 2018.	

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?

If the exact binomial test is not appropriate, what alternative method may be appropriate given the limitations described in the attachment?

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

If a 90% confidence level is not appropriate, what alternative confidence level may be appropriate given the limitations described in the attachment?

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 Pollutant (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: $r \leq p_2 = 0.20$	H_O: $r > p_1 = 0.10$, H_A: $r \leq p_2 = 0.25$

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.

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 December 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)**:**

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

a) chronic aquatic life toxics criteria? For larger samples sizes, >20, yes

b) conventional pollutant criteria? For larger samples sizes, >20, yes

If the exact binomial test is not appropriate, what alternative method may be appropriate given the limitations described in the attachment? For smaller sample sizes, the Exact Test makes sense, perhaps with the caveat that larger samples sizes are desired and would make for a more informed decision and should be sought out.

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

If a 90% confidence level is not appropriate, what alternative confidence level may be appropriate given the limitations described in the attachment?

80% is more appropriate confidence level given a) the precision observed from laborartories, b) given the various media types likely to be considered. As the matrix being analysed becomes more complex- turbid waters, waters with rich biological life, sediments, tissues- the reality of achieving 90% consistency in the reported value diminishes quickly. Many analytical requirements are in the +/- 10-30% range, depending upon the media and expected concentration range. While this +/- range is not to be confused with a given confidence intervals, it does not seem logical to expect a high confidence in an decision when the underlying data may not be produced with that level of confidence.

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 Pollutant (chronic criteria)	**Conventional Pollutants**
Listing	**H_O: $r \leq p_1 = 0.05$,** **H_A: $r > p_2 = 0.$~~20~~15**	**H_O: $r \leq p_1 = 0.10$,** **H_A: $r > p_2 = 0.$~~25~~20**
Delisting	**H_O: $r > p_1 = 0.05$,** **H_A: $r \leq p_2 = 0.$~~20~~15**	**H_O: $r > p_1 = 0.10$,** **H_A: $r \leq p_2 = 0.$~~25~~20**

DEQ's proposed null hypotheses and critical exceedance rates seem valid and in line with EPA recommendations and practices in other states. DEQ could more fully describe its rationale for the values selected, however. It appears that the Florida Department of Environmental Protection, for example, attempted to estimate sources and magnitude of variability in assessment data to help support their selection of critical exceedance rates. It is not clear that Oregon has the same type of analysis. Such an evaluation could show, for example, that the variability in toxicant measurements (e.g., as shown in split or duplicate samples or from other aspects of data quality assessment) is the same or higher than for conventional pollutants, leading to critical exceedance rates that are different than DEQ is currently proposing to use. The effect size of 15% seems to be supported by EPA guidance and use in other states. The approach of using different critical exceedance rates in the same hypothesis test for Ho and Ha also seems useful.

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?

Roughly, yes. Suggest slightly lower alternate hypothesis thresholds of 0.15 and 0.20 frequencies.

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

It does not require balancing. I suppose that could easily depend upon other considerations with respetect to relevant legislation or the risks of a given pollutant or other considerations. But, all elese being equal, yes, a balanced type II error rate for listing and delisting makes sense.

State of Oregon Department of Environmental Quality

Peer Review Materials

Date of Request: December 21, 2017

SCIENTIFIC PEER REVIEW: SOLICITATION REQUEST FORM

Reviewer Information	
Reviewer Name: Dr. Yangdong Pan	**Title:** Professor of Environmental Science & Management
Email Address:	**Contact Phone #:**
Employer: Portland State University	**Employer Category:** (federal agency, state agency, academic, professional organization/consultant)
Subject Matter: Exact Binomial Assessment Methodology **Timeline for Review Completion:** Reviews should be completed and returned electronically to DEQ by January 29, 2018.	

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?

If the exact binomial test is not appropriate, what alternative method may be appropriate given the limitations described in the attachment?

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

If a 90% confidence level is not appropriate, what alternative confidence level may be appropriate given the limitations described in the attachment?

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 Pollutant (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: $r \leq p_2 = 0.20$	H_O: $r > p_1 = 0.10$, H_A: $r \leq p_2 = 0.25$

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.

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 December 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)**:**

Using the statistical hypothesis testing methods such as the binomial test for watershed water quality violation listing is a step forward to improve DEQ's water quality assessment program, compared to the raw score method. The exact binomial test is one of the textbook hypothesis testing methods for binary outcomes and has been applied to assess water quality violation. My comments mainly focus on a few potential issues:

• The validity of a statistical hypothesis testing is largely depending on how well the population of interest is adequately sampled. DEQ proposes to use the exact binomial test to analyze the water quality monitoring data and detect violations. More importantly, the statistical method should also be used to guide the optimal sampling program designing. Decoupling these two may potentially give the public an impression that the listing or delisting is scientifically defensible simply because a well-established statistical method is used. In this case, a watershed is a target population. Water samples collected weekly, monthly or annually in the watershed are used to assess if the water quality standards are violated for the watershed. Like other classic statistical methods, the exact binomial test requires random and independent sampling of the population. It is not clear to me (largely due to my ignorance of the DEQ's water quality monitoring program) if these routine minoring samples are selected in a simple random way or in a probability-based and spatially balanced way (Stevens and Olsen 2004) and how DEQ assess if the sample is representative of the watershed. It is also not clear if the water samples collected are spatially and temporally independent. 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. In short, the sampling design should be considered as an integral part of the assessment of the validity of the statistical hypothesis testing.

• Statistical hypothesis testing using the frequentist approach has been questioned (Wasserstein and Lazar 2016). For instance, it has been pointed that the null hypothesis is not directly tested (Reckhow 1990). A test statistic such as t value or chi-square value is calculated based on the data collected and then the calculated test statistic is compared with the sampling distribution of the test statistic which is

generated under the assumption that the null hypothesis is true and the same study is repeated many times. The null hypothesis is rejected only if the observed test statistic based on the data is too extreme (defined by a significance level, typically 5%) compared to the sampling distribution of the test statistic. In short, the frequentist-based hypothesis testing is testing the observed data under the null hypothesis (P(data)|Ho). In this case, DEQ is using the hypothesis testing to determine if the water quality standard is violated and thus may be more interested in directly testing the hypothesis using the available data (P(Ho)|data). Reckhow (1990) articulated the above issue, particularly in non-replicated studies such as this case, and proposed Bayesian analysis as an alternative.

• I am not the expert on Bayesian methods but the Bayesian concept resonates with me well as an applied ecologist. Each watershed, as an ecosystem, varies substantially in terms of size, drainage network, geology, vegetation, land-use, etc. and to make it worse, the sampling efforts are often very limited due to the legitimate reasons which are well described in the White Paper. On the other hand, watershed managers may have a wealthy amount of qualitative information on the watershed, which can be very valuable in terms of the watershed water quality assessment. The 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. The water quality monitoring is a long-term effort and as time goes on, more data will be available. Again the Bayesian approach provides a way to update the model with newly available information.

• It seems to me that managing type II error is more costly and challenging than type I error in the watershed water quality assessment. A watershed may be falsely listed (type I error). It seems to me that if a follow-up monitoring program targeting on a location where the alleged violation occurs may provide more relevant data to address the error. On the other hand, due to several factors such as small sample size, heterogeneity of the watershed, and relatively small effect size (e.g., chronic effects), the chance to commit type II error may be high and may vary from one watershed to another. For the purpose of protecting water resources, it will be much more costly 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.

Cited references:

Reckhow, K. 1990. Bayesian Inference in Non-Replicated Ecological Studies. Ecology 71: 2053-2059.

Stevens Jr, D.L. and A. R. Olsen 2004. Spatially balanced sampling of natural resources. Journal of the American Statistical Association 99: 262-278.

Wasserstein, R. L. and N. A. Lazar 2016. The ASA's statement on p-values: context, process, and purpose. The American Statistician 70: 129-133.