Draft Appendix J. Evaluating Laboratory Performance with the Chronic *Ceriodaphnia dubia* Reproduction Toxicity Test

December 24, 2019

J.1 Overview and Key Findings

This appendix compiles and discusses recent research and information on how laboratory performance affects the pass or fail result of the chronic *Ceriodaphnia dubia* reproduction toxicity test when using the Test of Significant Toxicity (TST) and No Observed Effect Concentration (NOEC) statistical approaches. The goal of this appendix is to provide additional clarity and analysis of these statistical approaches and how laboratory performance, in terms of within-laboratory variability (precision), is key in ensuring the statistical power of the TST.

This appendix includes (1) key findings, (2) a summary of key statistical concepts, (3) an analysis of laboratory performance and its effect on the false positive probabilities when using the NOEC and the TST, (4) an assessment of the occurrences of fails at or below the 10 percent effect of permit compliance data, (5) and an analysis of probabilities of having an effluent limitation violation and being required to conduct a toxicity reduction evaluation (TRE) based on the probability of a fail at or below 10 percent effect.

This appendix focuses on chronic *C. dubia* reproduction toxicity tests for non-stormwater National Pollution Discharge Elimination System (NPDES) discharges. Therefore, samples of concern are effluent collected at the instream waste concentration (IWC). For additional discussion on statistical analysis, see Section 5.3 of this Staff Report.

Key Findings:

- When within-test variability is low and the percent effect is low, the NOEC is more likely to declare a sample toxic than the TST. When within-test variability is high and the percent effect is greater than or equal to 25 percent, the NOEC is less likely to declare a sample toxic, while the TST will always declare the sample toxic.

- Fox et al. 2019 examined data from 2012 to 2019 from a subset of California laboratories. Four of six laboratories had low within-test variability and, therefore, can attain the acceptable false positive probability of five percent using 10 test replicates (N=10). If the number of replicates were increased to 20 (N=20), then five of the six laboratories would meet the acceptable false positive probability.

- State Water Board staff examined more recent data from 2017 to 2019 from a subset of California laboratories. Three of four laboratories had low within-test variability and can attain the acceptable probability of a fail at or below 10 percent effect of five percent using 10 replicates. If the number of replicates were increased to 20, then all 4 laboratories would meet the acceptable probability.

- The TST statistical approach incentivizes laboratories to produce more precise data and increase statistical power. The Los Angeles County Sanitation District's San Jose Creek Water Quality Laboratory's (LACSD Municipal Laboratory) test performance improved when they began using the TST statistical approach. This was noted by Fox et al. 2019, as well as

independent analyses by the State Water Board staff. State Water Board staff also analyzed the test performance of a commercial laboratory that uses the TST statistical approach and concluded that their precision consistently improved from 2017 to 2019.

- The TST statistical approach is less likely than the NOEC statistical approach to identify a sample as toxic when biological effects are negligible (at or below a 10 percent effect) and will always identify a sample as toxic when percent effect is at or above a 25 percent effect level. Of the 984 California laboratory test results reviewed, there were no results of a fail when the percent effect was 10 percent or less, and no results of a pass when the percent effect was 25 percent or greater.

- The draft Toxicity Provisions state that more than one TST test fail in a calendar month is a median monthly effluent limitation (MMEL) violation, and two violations in a month or in two consecutive months will result in a requirement to conduct a TRE. The probability is very low of determining a single MMEL violation based on TST fails with a percent effect at or below 10. The probability of being required to conduct a TRE based on TST fails with a percent effect at or below 10 is even lower.

J.2 Relevant Statistical Concepts

This section describes relevant statistical concepts to inform subsequent sections.

The *true mean* is the mean for a theoretical statistical population of results from indefinite repetition of toxicity tests on the same control water and effluent sample. In contrast, the mean for the biological measure for a single toxicity test is referred to as the *sample mean*. (U.S. EPA 2010).

The *percent effect* (PE), or the mean percent effect, for a chronic *C. dubia* reproduction toxicity test is the difference between the control mean and the IWC treatment (sample) mean divided by the control mean. Restated, it is the difference between the mean number of neonates in the control replicates and the mean number of neonates in the IWC sample replicates, divided by the control mean. The percent effect does not reflect the amount of variability among replicates in a treatment. The TST statistical formula incorporates the measure of variability in determination of the test result. Figure J-1 illustrates how within-test variability is a determining factor in the TST result in relation to percent effect (Dr. Jerry Diamond, Personal Communication 2019).

Within-test (intra-test) variability is the variability in test organism response within a concentration averaged across all concentrations of the test material in a single test (U.S. EPA 2000).

Within-laboratory (intra-laboratory) variability is the variability that is measured when tests are conducted using specific methods under reasonably constant conditions in the same laboratory. Within-laboratory variability, as used in this document, includes within-test variability.

The *coefficient of variation* (CV) measures the relative variation of a data set. It is defined as the standard deviation divided by the mean and is sometimes known as the relative standard deviation. A lower CV value indicates lower within-test variability in the number of neonates produced in each individual replicate, compared to the mean. For the TST, the CV can be determined for both the control and sample (IWC) treatments. Often, the control CV data from a number of tests is used to assess within-laboratory variability over time.

The following terms are often used interchangeably:

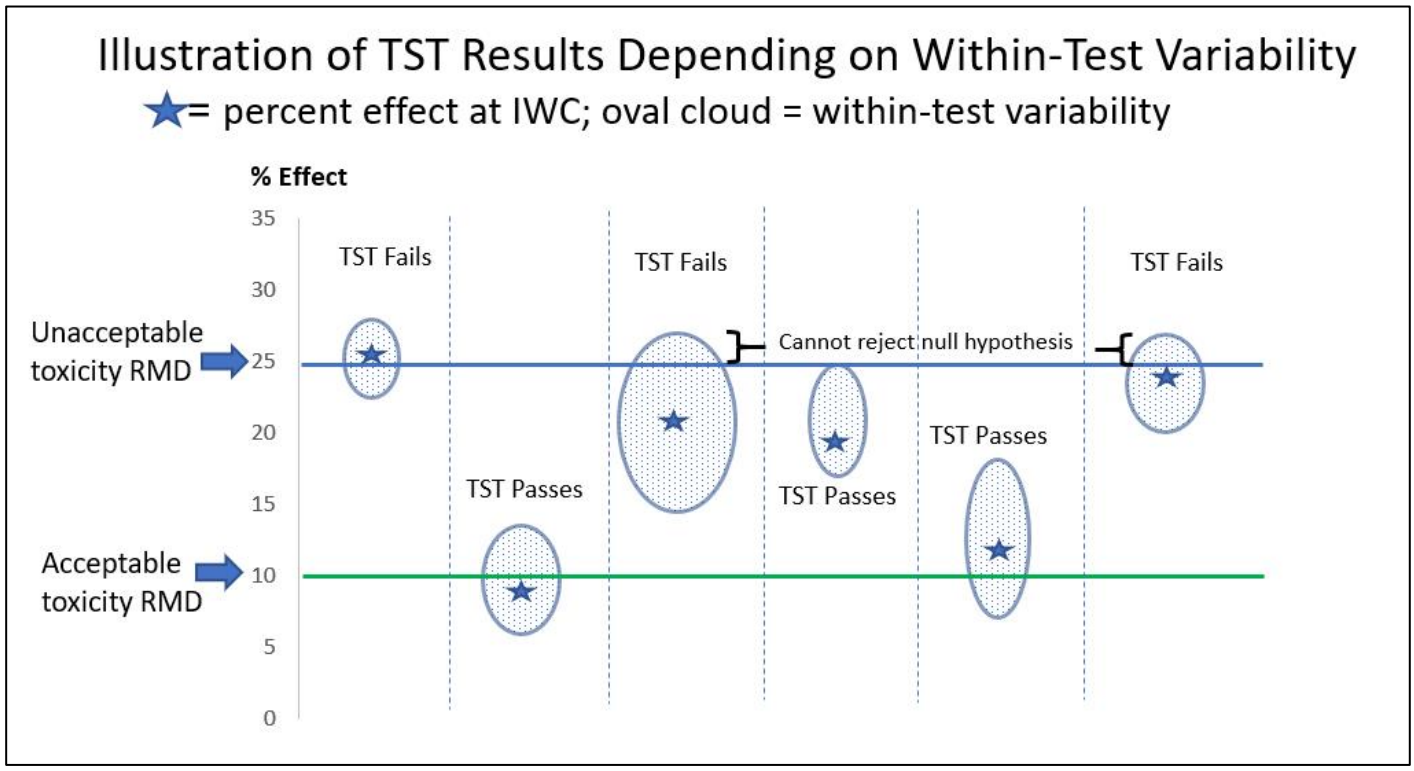
lower control coefficient of variation (CV) = lower within-test variability = higher precision

higher control coefficient of variation (CV) = higher within-test variability = lower test precision

A *false positive* is when the IWC sample is declared toxic (fail) but the sample is in fact not toxic. In the TST statistical approach, the false positive probability is the probability of a fail occurring when the percent effect is at 10 percent or less. No one effluent test can be called a "false positive" because of the variability of the data around the mean in the treatment and the control. The only true false positive is one where the sample is known to be absolutely non-toxic and the test results in a fail at or below the 10 percent effect.

A *false negative* is when the IWC sample is declared not toxic (pass) but the sample is in fact toxic. In the TST statistical approach, the false negative probability is denoted as alpha (α), and applies when the percent effect at the IWC is greater than or equal to 25 percent for a given test.

Figure J-1. Illustration of TST Results Depending on Within-Test Variability



J.3. False Positive and Negative Probabilities and Laboratory Performance

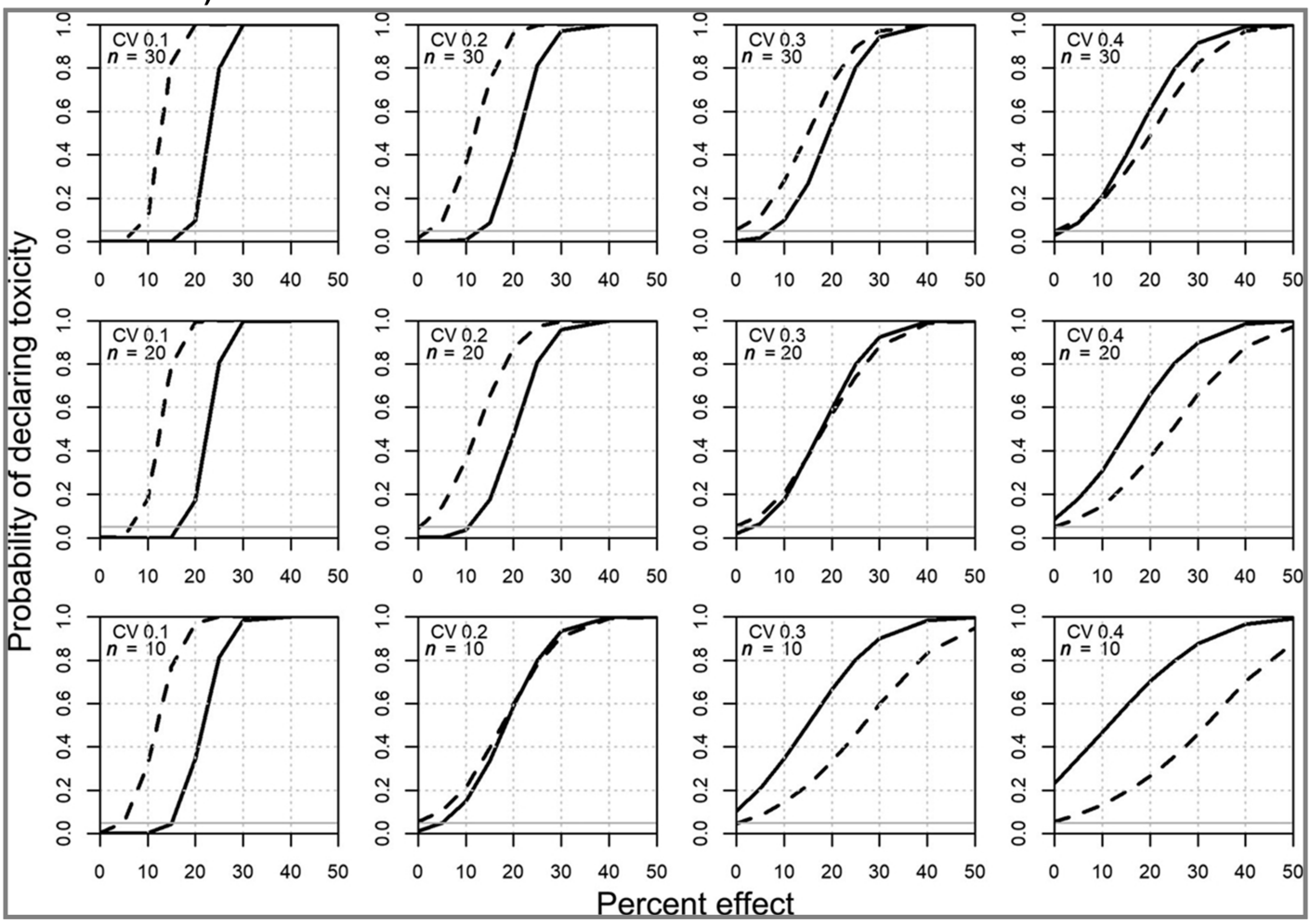
In March 2019, the Environmental Toxicity and Chemistry journal published a peer-reviewed article titled *Comparison of False-Positive Rates of 2 Hypothesis-Test Approaches in Relation to Laboratory Toxicity Test Performance* by Dr. Fox, Dr. Denton, Dr. Diamond, and Ms. Stuber (Fox et al. 2019). The article compares the false positive and false negative rates of the NOEC and the TST to illustrate the effect of laboratory performance when using U.S. EPA's *C. dubia* reproduction toxicity test.

Probabilities of Declaring Toxicity Using the NOEC and TST Statistical Approaches

Figure J-2 presents the probability curves from Fox et al. 2019 paper. Precision is measured as the control CV. The columns from left to right show the probabilities of declaring a sample toxic with

decreasing precision. The rows from bottom to top to show probabilities of declaring a sample toxic as the number of replicates increases.

Figure J-2. Probabilities of Declaring a Sample Toxic When Using the NOEC and TST (From Fox et al. 2019)



From Fox et al 2019: Probability of declaring a sample toxic using the no-observed-effect concentration (NOEC) and test of significant toxicity (TST) based on simulating 10 000 whole-effluent toxicity tests at each of various percent effect parameter values (horizontal axis), 4 values of control coefficient of variation parameter, and 3 values for number of test replicates. Gray horizontal line shows probability of 0.05. Solid curves represent TST and broken curves, NOEC. CV= coefficient of variation.

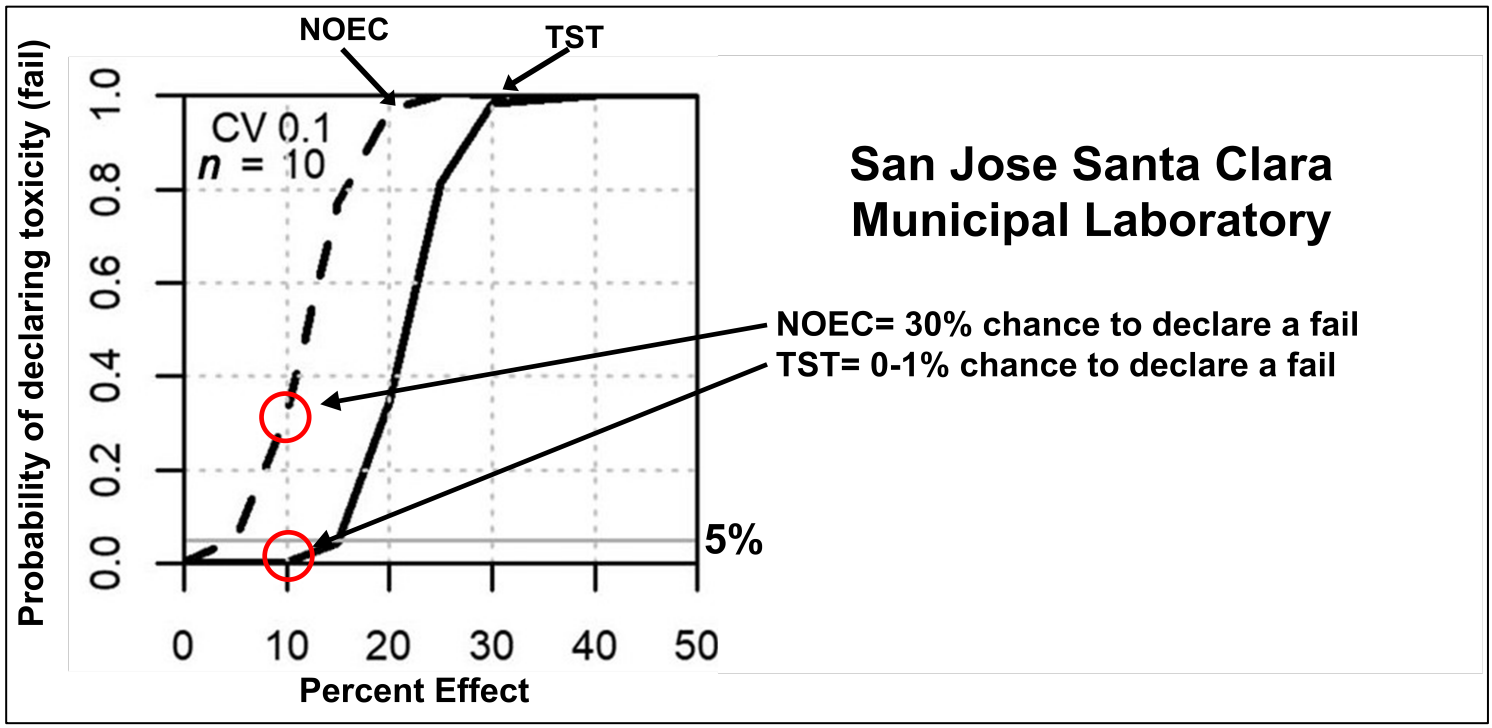
Comparison of NOEC and TST False Positive Error Rates

Fox et al. 2019 found that the TST statistical approach incentivizes laboratories to produce more precise data and increase statistical power. When within-test variability is low and the percent effect is low, the NOEC is more likely to declare a sample toxic than the TST. When within-test variability is high and the percent effect is high, the NOEC is less likely to declare a sample toxic than the TST.

When precision is high, the NOEC has a higher probability of declaring a fail when the percent effect is less than or equal to the 10 percent effect. A real-world example of high precision achieved in this range is the City of San Jose / Santa Clara Water Pollution Control Plant Laboratory (San Jose Santa Clara Municipal Laboratory), which, per Fox et al. 2019, had a median control CV of 0.11. Figure J-2 shows the probability curve for this CV and replicate number, which

was taken from the larger suite of probability curves in Figure J-2. Using the NOEC statistical approach (the dashed line), there is approximately a 30 percent chance of declaring the sample toxic when there is a 10 percent effect. Using the TST statistical approach (the solid line), there is less than one percent chance of declaring the sample toxic when there is a 10 percent effect.

Figure J-3. San Jose Santa Clara Municipal Laboratory Probability Curves for NOEC and TST (From Fox et al. 2019)



The San Jose Santa Clara Municipal Laboratory has low within-laboratory variability with 10 replicates, and they can attain the acceptable probability of a fail at or below 10 percent effect (five percent). For laboratories who have a higher median control CV, the replicate number of 10 may be inadequate to meet the acceptable false positive probability. Laboratories that currently have a median control CV approaching 0.2 would need to reduce within-laboratory variability and/or increase their number of replicates[1]. The additional replicate number needed are not required to be in multiples of 10. The number can be calculated based on the laboratory's median control CV (Fox et al. 2019).

California Laboratory Performance

In this section California laboratory performance data is presented from two different analyses. The first is from eight different laboratories that were presented in the Fox et al. 2019 paper. The second is from four different laboratories[2] analyzed by State Water Board staff to assess the ability of laboratories to attain the acceptable probability of a fail at or below 10 percent effect.

[1] Analysis using the TST is conducted comparing the laboratory control to the IWC treatment. Laboratories may choose to increase their number of replicates above 10 for the control and IWC, if desired.

[2] There are three unknown California laboratories (laboratories 1, 2, 3) and the same LACSD Municipal Laboratory analyzed by Fox et al. 2019. Because the names are not known for the three laboratories, there is the possibility that some or all of the three laboratories used in the State Water Board analysis might be some of the same laboratories in

Table J-1 summarizes the individual laboratory control CV information from Fox et al. 2019 for the eight California laboratories.

Table J-1. Chronic *C. dubia* Reproduction Control CV Values for Eight California Laboratories from Fox et al. 2019

Laboratory	Time Period	Number of Tests	Median Control CV	False Positive Probability met with 10 Replicates*	False Positive Probability met with 10 – 20 Replicates*
A-Commercial Lab	2012-2015	43	0.23	no	no
B-Commercial Lab	2012-2015	18	0.15	yes	yes
C-Commercial Lab	2012-2015	20	0.20	no	yes
D-2 LACSD Municipal Laboratory	2012-2015	57	0.10	yes	yes
E-Commercial Lab	2012-2015	22	0.11	yes	yes
F-San Jose Santa Clara Municipal Laboratory	2012-2015	20	0.11	yes	yes
D-1 LACSD Municipal Laboratory	Pre-2012 TST Test Drive	30	0.17	no	yes
G-Commercial Lab	Pre-2012 TST Test Drive	17	0.09	yes	yes
H-Commercial Lab	Pre-2012 TST Test Drive	17	0.10	yes	yes

* Based on Fox Probability Memo 2019**.** Probabilities were produced by R function *TST.pwr.fn2*, posted originally at https://figshare.com/articles/WET_Error_Rates_for_TST_NOEC_Supporting_Information/7122812 as a supplement to Fox et al. 2019. The function is included in the supporting document "Rfunctions-MMEL.R."

Prior to 2012

For both laboratories G and H, within-test variability prior to the *Effluent Stormwater and Ambient Toxicity Test Drive Analysis of the Test of Significant Toxicity* (TST Test Drive; SWRCB 2011) was low (median control CV less than or equal to 0.10). The actual test result data provided by these laboratories are part of the TST Test Drive analysis discussed later in this appendix. For the LACSD Municipal Laboratory, a replicate number between 10-20 would be needed to meet the acceptable probability of a fail at or below 10 percent effect (five percent).

Fox et al. 2019. The data sets do not overlap temporally, therefore the analysis of the data sets are at a minimum, unique for time period.

From 2012 to 2015

For the post TST Test Drive results, four of six California laboratories have low within-test variability and can attain the acceptable probability of a fail at or below 10 percent effect (five percent) with 10 replicates. If the number of replicates increases between 10 and 20, then five of the six laboratories would meet the acceptable probability of a fail at or below 10 percent effect with their current median control CVs.

From 2017 to 2019

According to data from the three commercial laboratories and the LACSD Municipal Laboratory from the 2017 through 2019 time period, three of the four laboratories have low within-test variability and can attain the acceptable probability of a fail at or below 10 percent effect with 10 replicates and their current median control CVs. For Commercial Laboratory #3, within-test variability has decreased each year (from a CV of 0.24 in 2017 to a CV of 0.16 in 2019), and 11 replicates would be needed to meet the acceptable probability of a fail at or below 10 percent effect.

Table J-2. Chronic *C. dubia* Reproduction Control CV Values for Four California Laboratories from Submitted Data & CIWQS

Laboratory	Time Period	Number of Tests	Median Control CV	False Positive Probability met with 10 Replicates[1]	False Positive Probability met with 10 – 20 Replicates
Commercial Laboratory #1	2018-19	75	0.08	yes	yes
Commercial Laboratory #2	2019	75	0.12	yes	yes
Commercial Laboratory #3	2019	100	0.16	no	yes
LACSD Municipal Laboratory	2017-18	203	0.13	yes	yes

Improvements in Laboratory Performance

State Water Board staff analyzed control CV data to compare within-laboratory variability before and after the TST statistical approach was required in LACSD's wastewater discharge permits. The Los Angeles Regional Water Board began to incorporate the TST statistical approach into NPDES permits in May 2014. Table J-3 contains data for two laboratories who conduct the *C. dubia* test with the TST statistical analysis; the LACSD Municipal Laboratory and Commercial Laboratory #3. Figure J-4 presents the entire data set of the LACSD Municipal Laboratory's minimum, median, and maximum control CV's from 835 chronic *C. dubia* reproduction toxicity tests run between 2010 and 2018.

LACSD Municipal Laboratory

Both the Fox et al. 2019 researchers and State Water Board staff found LACSD Municipal Laboratory improved their laboratory performance after 2012, as demonstrated by reductions in the control CVs and within-test variability. At their most recent median control CV, LACSD Municipal Laboratory would need 10 replicates to attain the acceptable false positive probability. The district

is currently running 20 replicates at the control and IWC. By running the additional replicates, the probability of declaring a sample toxic is less than one percent when the percent effect is 10 percent or less. This is well below the acceptable probability of a fail at or below 10 percent effect (five percent).

Commercial Laboratory #3

In Table J-3, Commercial Laboratory #3 in 2017 had the highest median control CV of the four laboratories. This laboratory is known to conduct a portion of their tests using the TST statistical approach for compliance. Looking at data from three years, there has been a consistent improvement of performance at Commercial Laboratory #3. For the 2019 median control CV data, this laboratory could meet the acceptable probability of a fail at or below 10 percent effect (five percent) with a replicate number of 11.

Table J-3. Chronic *C. dubia* Reproduction Control CV Values for LACSD Municipal Laboratory and Commercial Laboratory #3 Over Time

Laboratory	Time Period	Number of Tests	Median Control CV	Acceptable False positive Probability met at N=10	Acceptable False positive Probability met at N= 10 – 20
LACSD Municipal Laboratory	Pre-2012 TST Test Drive	30	0.17	no	yes
LACSD Municipal Laboratory	2012-2015	57	0.10	yes	yes
LACSD Municipal Laboratory	2017-2018	203	0.13	yes	yes
Commercial Laboratory #3	2017	93	0.24	no	yes
Commercial Laboratory #3	2018	142	0.19	no	yes
Commercial Laboratory #3	2019	100	0.16	no	yes

California vs. National Laboratory Performance

Fox et al. 2019 also found that within-test variability of assessed California laboratories is comparable to national laboratories that were assessed in U.S. EPA's TST Implementation Document (U.S. EPA 2010). Table J-4 compares the percentiles of *C. dubia* reproduction toxicity test control CV values between the national TST Technical Document and the eight California laboratories assessed in Fox et al. 2019. The "Percentile" column represents the percentage of tests (in the specified data set) which had a control CV less than the specified value. For example: of the 244 toxicity tests examined in Fox et al. 2019, 90% of the tests had a control CV less than 0.332.

The median CV value (i.e., the 50th percentile) for California laboratories assessed by Fox et al. 2019 is 0.147. This demonstrates that California laboratories' performance is consistent with other laboratories across the nation and are able to successfully conduct chronic *C. dubia* reproduction toxicity tests with low within-in test variability.

Figure J-4. Minimum, Median, and Maximum Control CV Values for the LACSD Municipal Laboratory from 2010 through 2018

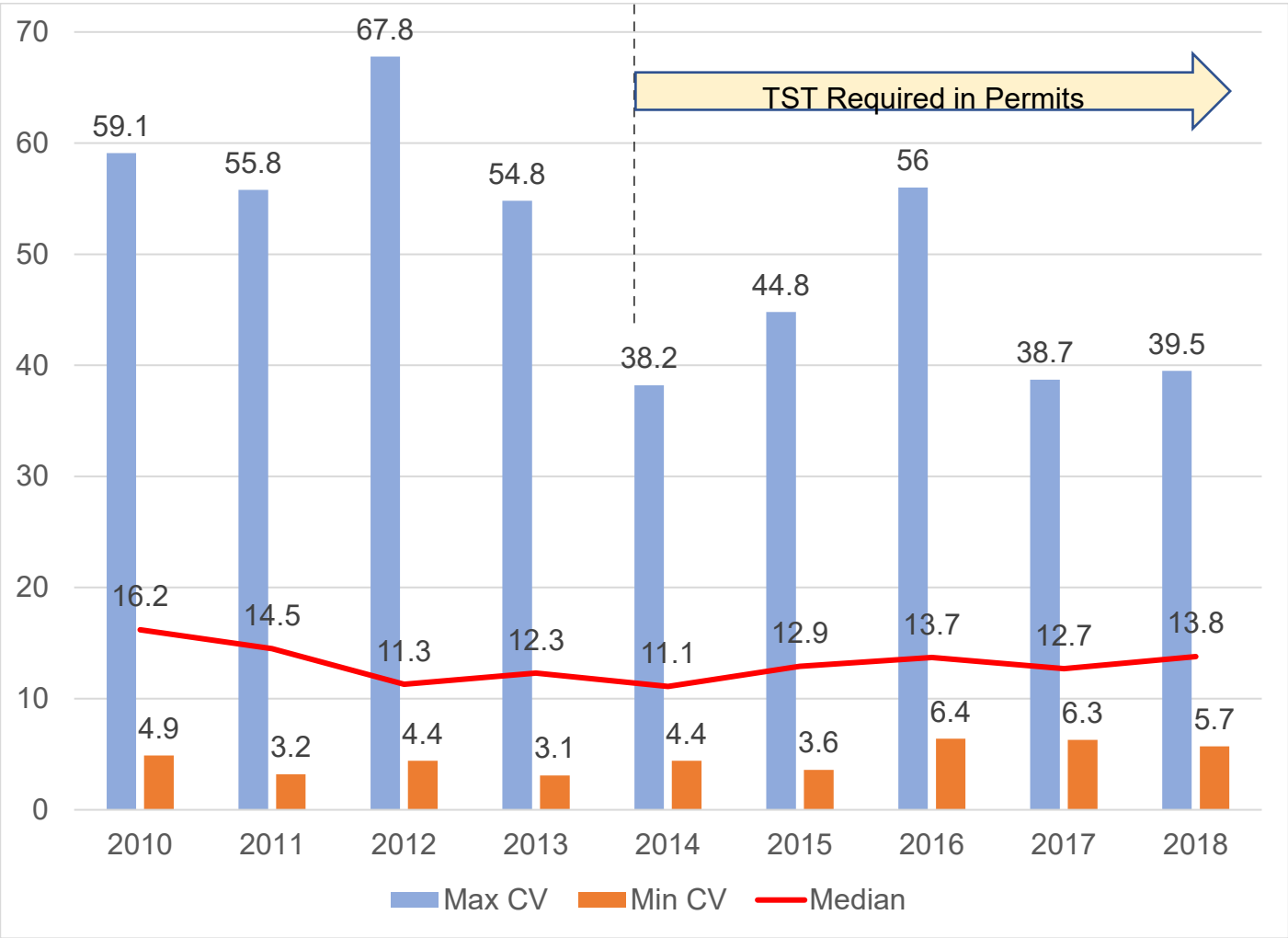


Table J-4. Comparison of Percentiles of *C. dubia* Control CVs between the National Study (U.S. EPA 2010) and Fox et al. 2019

Percentile	Control CV's from U.S. EPA 2010 TST Technical Document	Control CV's From Fox et al. 2019 Study
0%	—	0.036
10%	0.08	0.076
25%	0.1	0.097
50%	0.15	0.147
75%	0.24	0.244
90%	0.35	0.332
100%	—	0.568
Number of Tests	792	244
Number of Laboratories	44	8

J.4. Permit Compliance Data

The appendix discussion to this point has focused on how laboratory performance affects meeting the acceptable false positive probability of a fail at or below 10 percent effect. State Water Board staff analyzed six data sets of actual test results using TST to evaluate how often a sample was declared toxic when the percent effect was 10 percent or less, and how often a sample was declared not toxic when the percent effect was 25 percent or greater. The test data were analyzed with the TST statistical approach, except for the TST Test Drive data, which has both the NOEC and TST results. The individual facilities were not identified in most of the data sets.

Under the NOEC, historic data has shown that high within-test variability results in a higher number of passes when the percent effect is greater than or equal to 25 percent than the TST (SWRCB 2011; Diamond et al. 2013). Conversely, the NOEC will result in a fail at a percent effect less than or equal to 10 percent more often when within-test variability is low (Diamond et al. 2013; Fox et al. 2019).

The following six data sets were analyzed as these data were readily available and had already been compiled for other purposes:

- TST Test Drive - NPDES facilities only
- The County of Los Angeles
- The City of Los Angeles
- City of Simi Valley
- City of San Jose Santa Clara Municipal Laboratory
- City and County of Honolulu

The staff evaluation found that the TST statistical approach is less likely to identify a sample as toxic when biological effects are negligible (at or below a 10 percent effect) and more likely to identify a sample as toxic when effects are biologically significant (at or above a 25 percent effect) than the NOEC statistical approach. Of the 984 test results reviewed from California laboratories, there were no results of a fail when the percent effect was 10 percent or less, and no results of a pass when the percent effect was 25 percent or greater.

Comparison of NOEC and TST Results Using TST Test Drive Data

Staff queried data compiled for the TST Test Drive for the six NPDES wastewater facilities and analyzed 209 chronic *C. dubia* reproduction toxicity test results using both the NOEC and TST statistical approaches. The TST Test Drive is described in Section 5.3.1 of this Staff Report.

Figure J-5 shows the toxicity data evaluated using the NOEC, and Figure J-6 shows the same data evaluated using the TST. Both figures highlight the number of times tests resulted in a fail (i.e., a determination of toxicity) or a pass (i.e., a determination of no toxicity) and the calculated percent effect.

When using the NOEC statistical approach, there were three results when there was a fail when the percent effect was 10 percent or less. There were five where there was a fail between 10 and 25 percent effect. There were five results when there was a pass when the percent effect was 25 percent or greater. When using the TST statistical approach, there were no results of a fail when the percent effect was 10 percent or less. There were 13 results of a fail in the 10-25 percent effect range, and no results of a pass when the percent effect was 25 percent or greater.

Figure J-5. TST Test Drive Results for NPDES Facilities Using the NOEC Analysis

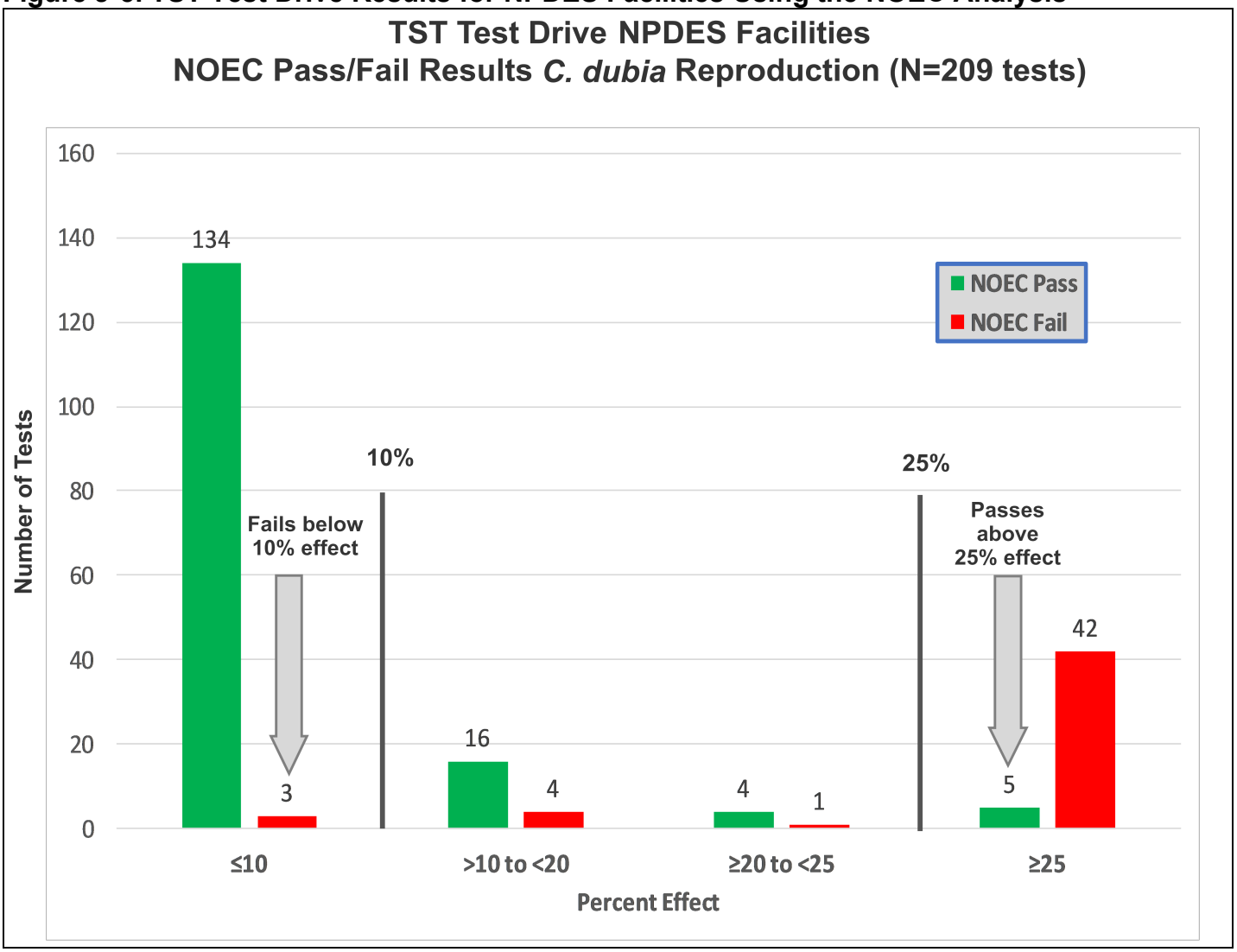
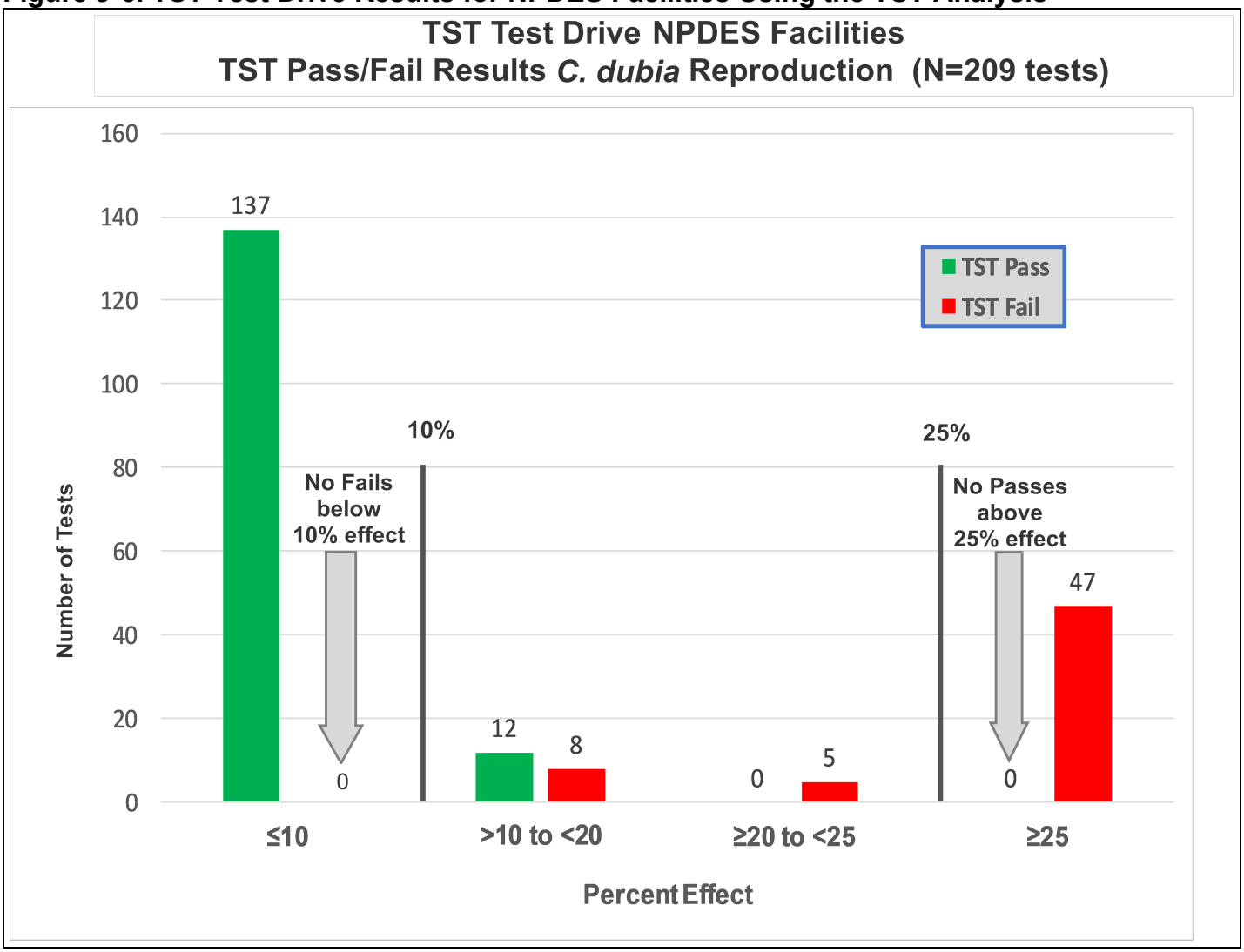


Figure J-6. TST Test Drive Results for NPDES Facilities Using the TST Analysis



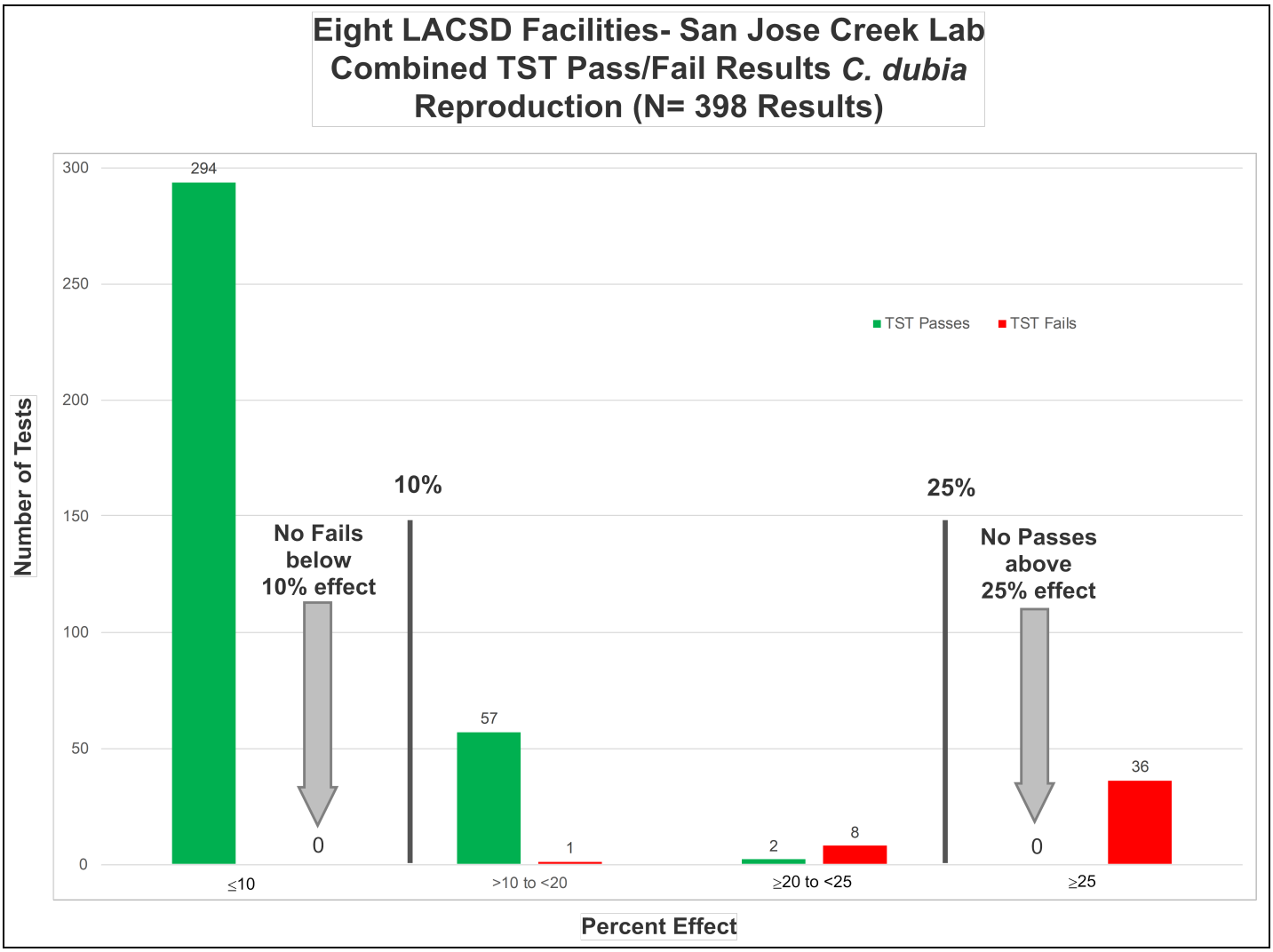
TST Passes and Fails by Percent Effect for LACSD Municipal Laboratory TST Results

The discussion above of LACSD Municipal Laboratory performance shows strong statistical power for meeting the RMD and low probability of declaring a sample toxic when the percent effect is equal to or less than 10 percent.

State Water Board staff analyzed 398 chronic *C. dubia* reproduction toxicity test results generated between 2015 and 2019 for eight LACSD facilities. The majority of the tests were conducted by the LACSD Municipal Laboratory. Figure J-7 shows the toxicity data evaluated using the TST statistical approach and highlights the number of times tests resulted in a fail or a pass and the associated percent effect.

There were no results of a fail when the percent effect was 10 percent or less. There was one fail in the 10 to 20 percent range (at 19.4 percent effect), and eight fails between 20 and 25 percent effect. There were no results of a pass when the percent effect was 25 percent or greater.

Figure J-7. LACSD Municipal Laboratory TST Test Results

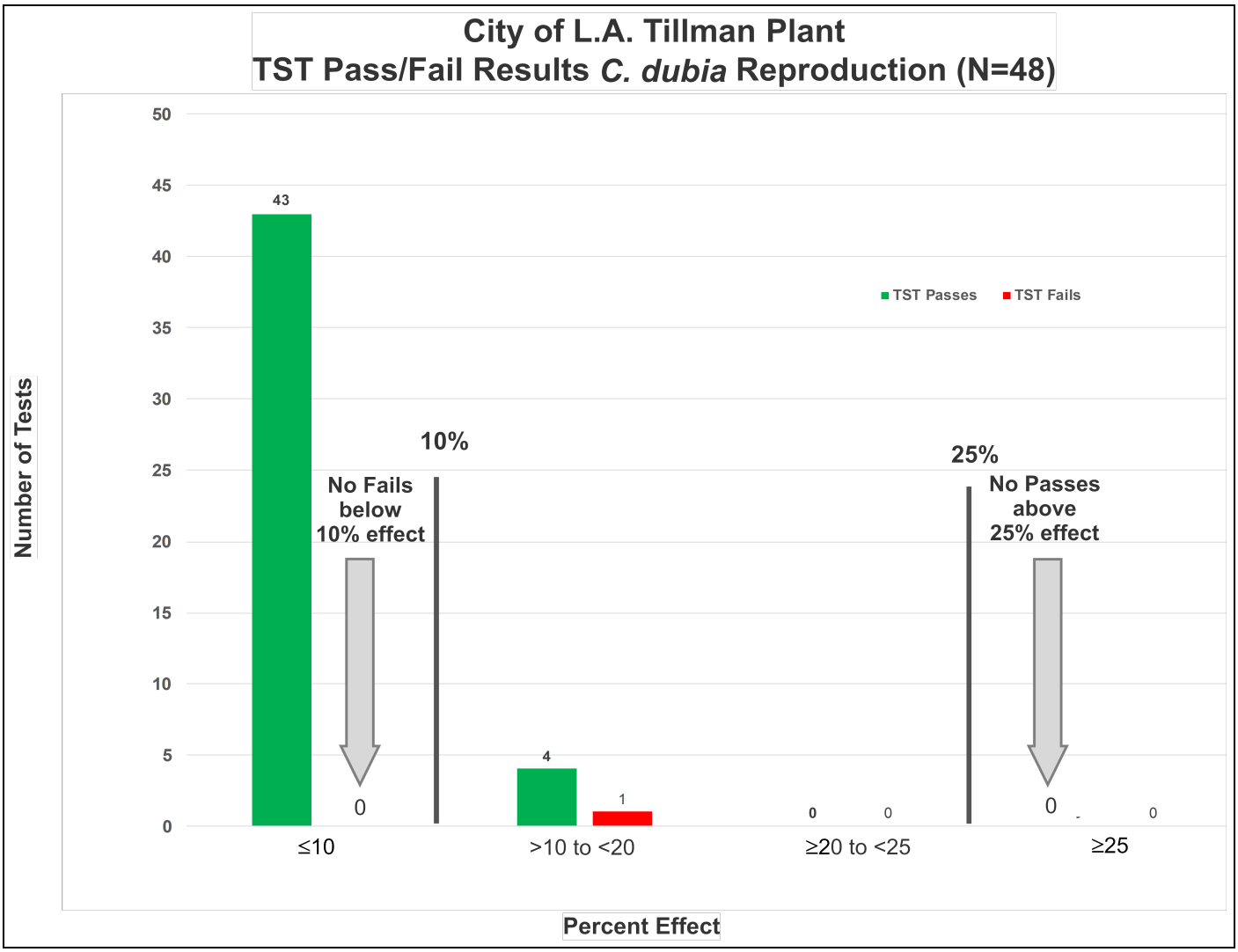


The absence of fails at or below the 10 percent effect and passes above the 25 percent effect is consistent with the probability curves in Figure J-2 when applied to the median control CV for LACSD Municipal Laboratory. LACSD Municipal Laboratory uses 20 replicates for testing the control water and the IWC. Given their most recent median control CV of 0.13, the probability of declaring a sample a fail is less than one percent when the percent effect is 10 percent or less. Additionally, the eight fails between 20 and 25 percent effect are consistent with the probability curve as well. As the percent effect approaches 25 percent, the probability of declaring a sample toxic increases.

TST Passes and Fails by Percent Effect for the City of Los Angeles Test Results

Staff analyzed 48 chronic *C. dubia* reproduction toxicity test results from the City of Los Angeles' Donald C. Tillman Water Reclamation Plant. Figure J-8 shows the toxicity data evaluated using the TST statistical approach and highlights the number of times tests resulted in a fail or a pass and the associated percent effect. There were no results of a fail when the percent effect was 10 percent or less. There were no results of a pass when the percent effect was 25 percent or greater.

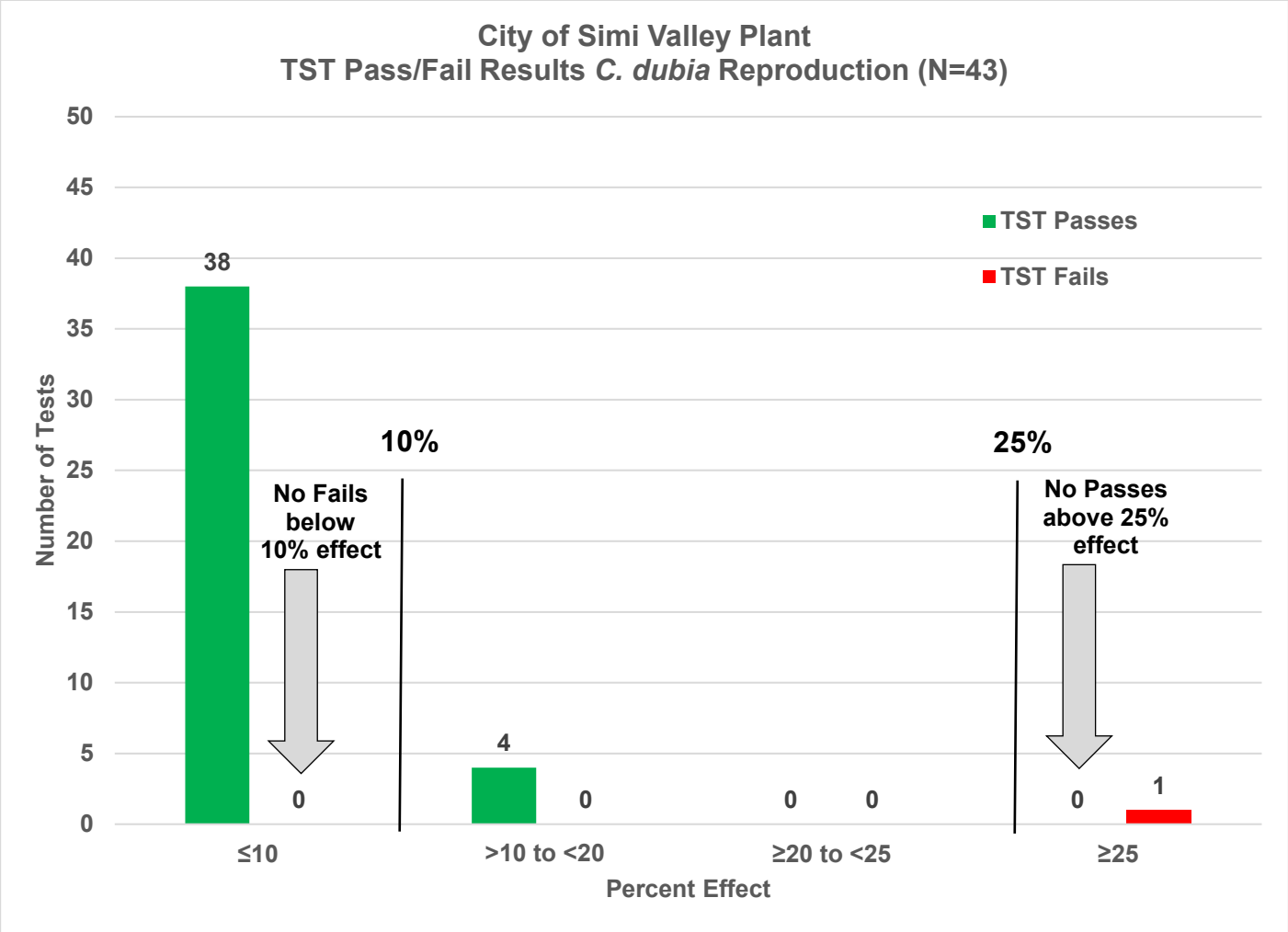
Figure J-8. City of L.A. Tillman Plant TST Test Results



TST Passes and Fails by Percent Effect for City of Simi Valley Test Results

Staff analyzed 43 chronic *C. dubia* reproduction toxicity test results from the City of Simi Valley's Wastewater Treatment Plant. Figure J-9 shows the toxicity data evaluated using the TST statistical approach and highlights the number of times tests resulted in a fail or a pass and the associated percent effect. There were no results of a fail when the percent effect was 10 percent or less. There were no results of a pass when the percent effect was 25 percent or greater.

Figure J-9. City of L.A. Simi Valley Plant TST Test Results



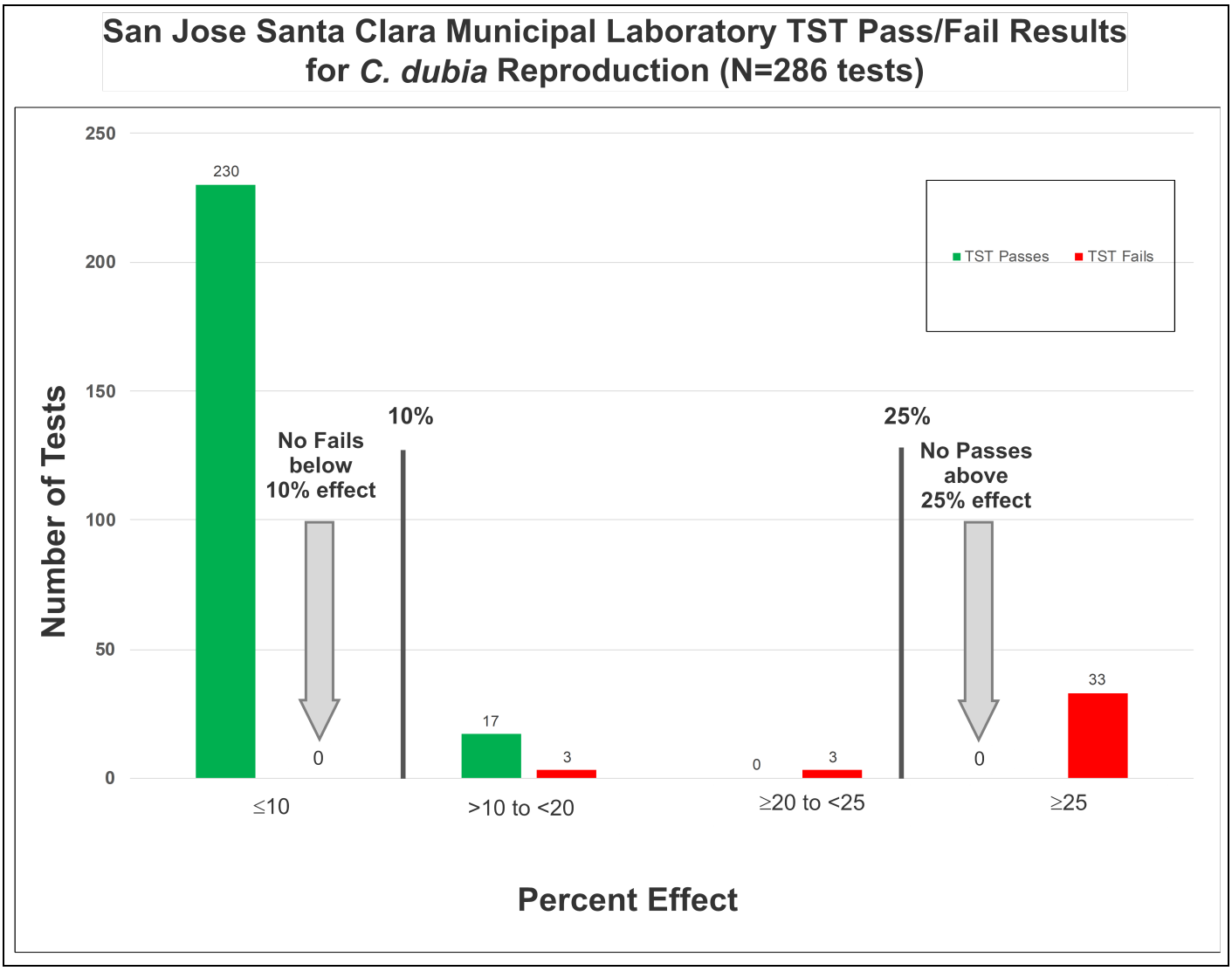
TST Passes and Fails by Percent Effect for the San Jose Santa Clara Municipal Laboratory Test Results

State Water Board staff analyzed 286 chronic *C. dubia* reproduction toxicity test results conducted between 1996 and 2016 by the San Jose Santa Clara Municipal Laboratory. For some test results, the data show a mean percent effect as a negative value, meaning the mean number of neonates in the IWC sample was higher than the mean number of neonates in the control water. Other data lacked percent effect values but were labelled as stimulatory (i.e. the effluent sample stimulated higher reproduction in the IWC treatment compared to the control). The laboratory grouped data from both of these categories with tests that had a percent effect from zero to 10 percent.[3] State Water Board staff considered all these data to be at a less than or equal to 10 percent effect level. Figure J-10 shows the toxicity data evaluated using the TST statistical approach and highlights the number of times tests resulted in a fail or a pass and the associated percent effect.

[3] This data set reflects the changes over 20 years in laboratory data entry for aquatic toxicity tests. Researchers assessed 20 recent CV values from this laboratory and found it to have the highest precision of the laboratories studied (Fox et al. 2019).

There were no results of a fail when the percent effect was 10 percent or less. There were no results of a pass when the percent effect was 25 percent or greater.

Figure J-10. San Jose Santa Clara Municipal Laboratory TST Test Results



TST Passes and Fails by Percent Effect for the Hawaii TST Test Drive

The City and County of Honolulu conducted an internal data analysis similar to California's TST Test Drive (Vazquez 2012). The Hawaii TST Test Drive evaluated 255 chronic *C. dubia* reproduction toxicity test results conducted between 2007 and 2012 from four NPDES-permitted wastewater treatment plants in Hawaii. Data were assessed using both the NOEC and TST statistical approaches.

When using the TST statistical approach, there were no results of a fail when the percent effect was 10 percent or less, and there were four fails in the 15 to 25 percent effect range. When using the NOEC statistical approach, there were no fails in the 15 to 25 percent effect range. There were no results provided at a percent effect of 25 percent or greater.

Vazquez 2012 concluded on page 7 that:

> Toxic effects of effluents on C. *dubia* reproduction are difficult to detect with the NOEC approach because of the inherent within-test variability of this chronic WET test. The alternative TST procedure controls false negatives and identifies toxicity that may have potential adverse environmental effects.

Vazquez 2012 also concluded on page 7 that:

> The failures [in the 15 to 25 percent effect range] declared by TST in this study were very rare excursions caused by an episode of unusually poor C. *dubia* culture performance. While blocking by parentage minimizes within-test variability, the effect of limited fecundity or mortality of even a single organism may be remarkable. For this reason, there must be an extremely thorough oversight of laboratory protocols to ensure consistent organism vigor. In addition, increased replication in the control and in the sample at the IWC may be adopted to decrease variance.

For the four fails identified between a 15 and 25 percent effect using the TST statistical approach, the single test control CVs ranged from 0.28 to 0.36, indicating higher within-test variability. The City and County of Honolulu acknowledge this issue of precision in the conclusion above. By using the probability curve in Figure J-1 that corresponds to a CV of 0.3 and 10 replicates, it is more likely that use of the TST will declare toxicity at a 15 to 25 percent effect range than the NOEC statistical approach. However, the review of a single test control CV is not a complete analysis of toxicity test performance nor within-laboratory variability.

In December 2013, Hawaii's Department of Health adopted the TST statistical approach for assessing toxicity to evaluate the combined impact of all pollutants on aquatic organisms for Clean Water Act regulations (Hawaii State Department of Health 2015).

J.5. The Probability of a Violation and TRE Based on Laboratory Performance

Much of the analyses (SWRCB 2011, Vazquez 2012, Diamond 2013, Fox et al. 2019) and discussion above focuses on the probabilities and occurrences of a single test fail or pass. The Toxicity Provisions include MMELs designed to address possible effects of a discharge over a period of a calendar month. When a chronic or acute routine monitoring test results in a fail of the TST, the discharger would be required to conduct up to two MMEL compliance tests, initiated within the same calendar month. A violation of the MMEL occurs when two or more tests result in a fail in a calendar month. A TRE would be required when there are two aquatic toxicity effluent limitation violations (MMEL or MDEL) in a calendar month or in consecutive calendar months.

At the request of State Water Board staff, Dr. John Fox prepared a memorandum titled *Probability of Failing TST and WET Maximum Monthly Effluent Limit* (Fox 2019). The memorandum includes statistical analyses of probabilities of TST fails when the percent effect is at 10 percent or less, the probabilities of these fails resulting in a violation of effluent limitations, and the probabilities of two violations based on these fails resulting in the requirement to conduct a TRE. Dr. Fox found that, by requiring two out of three TST test fails before receiving a MMEL violation, the probabilities of receiving a MMEL violation based on a fail at or below the 10 percent effect are low. As a result, the probabilities of violations occurring based on a fail at or below the 10 percent effect that ultimately require a TRE are even lower.

The probability of a maximum daily effluent limitation (MDEL) violation occurring in combination with a MMEL violation, and the subsequent TRE requirement, is not applicable in this discussion.

The MDEL is violated when the test results in a fail of the TST and a 50 percent effect in the survival endpoint. At that percent effect level, the sample will always be declared toxic, independent of the laboratory precision.

Probability of Declaring Toxicity for a Single Test Based on Laboratory Precision

Table J-5 shows the probabilities of declaring a sample toxic at the different percent effects when using the TST statistical approach, as calculated by Dr. Fox. The probabilities depend on the within-test variability, which is expressed by the median control CV value over time, and the number of replicates used in the toxicity test.

In order to illustrate the probabilities associated with a MMEL violation and initiation of a TRE, a laboratory control CV of 0.15 was selected. This is the median CV value for the eight California laboratories assessed by Fox et al. 2019 and summarized in Table J-2. Additionally, the control CV of 0.15 is appropriate for this analysis as 0.15 is close to or higher than several median control CV values for the laboratory data summarized in Tables J-3 and J-4.

With a control CV of 0.15 and 10 replicates, there is a 4.8 percent probability of the TST statistical analysis resulting in a fail with a percent effect less than or equal to 10%. This is less than the five percent probability deemed acceptable. When the number of replicates increases to 20, there is a 0.3 percent probability of the TST statistical analysis resulting in a fail with a percent effect less than or equal to 10%.

Table J-5. Probability of Declaring Toxicity for Different Percent Effects Based on Laboratory Performance as Measured by the Median CV

Probability of failing TST for specified parameters percent effect and control CV, using 10 replicates	Percent Effect: 0%	Percent Effect: 10%	Percent Effect: 25%	Percent Effect: 50%
Control CV: 0.100	0.000	0.002	0.800	1.000
Control CV: 0.150	0.000	0.048	0.800	1.000
Control CV: 0.200	0.011	0.150	0.800	1.000
Control CV: 0.300	0.107	0.341	0.800	0.998
Control CV: 0.400	0.235	0.461	0.800	0.992
Probability of failing TST for specified parameters percent effect and control CV, using 20 replicates	**Percent Effect: 0%**	**Percent Effect: 10%**	**Percent Effect: 25%**	**Percent Effect: 50%**
Control CV: 0.100	0.0000	0.0000	0.8000	1.0000
Control CV: 0.150	0.0000	0.0030	0.8000	1.0000
Control CV: 0.200	0.0000	0.0340	0.8000	1.0000
Control CV: 0.300	0.0170	0.1740	0.8000	1.0000
Control CV: 0.400	0.0830	0.3110	0.8000	0.9990
Probability of failing TST for specified parameters percent effect and control CV, using 30 replicates	**Percent Effect: 0%**	**Percent Effect: 10%**	**Percent Effect: 25%**	**Percent Effect: 50%**
Control CV: 0.100	0.0000	0.0000	0.8000	1.0000
Control CV: 0.150	0.0000	0.0000	0.8000	1.0000
Control CV: 0.200	0.0000	0.0070	0.8000	1.0000
Control CV: 0.300	0.0030	0.0900	0.8000	1.0000
Control CV: 0.400	0.0300	0.2130	0.8000	1.0000

Probability of a MMEL Violation Based on the False Positive Rate

In Dr. Fox's simulation, he points out that for each of the up to three tests conducted in a row to determine MMEL compliance, laboratory control CV, percent effect, and number of replicates may vary. For this probability simulation, each of the tests' probabilities (p) of meeting the RMD are assumed to be the same (p1=p2=p3) and independent of each successive TST failure.

Table J-6 shows probability of a MMEL violation based on TST fails at or below 10 percent effect. Continuing with the illustration, for a control CV of 0.15 and 10 replicates, the probability (p) is 0.048 (less than five percent) of declaring the test a fail. With more than one test needed to result in an MMEL violation, there is only a 0.49 percent probability of a MMEL violation based on the false positive probability, which is 10 times less likely than the probability of a single test fail.

Table J-6. Probability of a MMEL Violation Based on TST Fails at or Below 10 Percent Effect

Probability of each test in simulation	p = Probability of a Fail ≤ 10 Percent Effect	Probability of a MMEL Violation
p1 = p2 = p3	0.05	0.0049
p1 = p2 = p3	0.10	0.0190
p1 = p2 = p3	0.15	0.0416
p1 = p2 = p3	0.20	0.0720
p1 = p2 = p3	0.25	0.1094

Probability of a MMEL Violation Based on the Probability of a Fail at or below 10 Percent Effect for Specific California Laboratories

The example above was based on the median CV value of 0.15 for multiple California laboratories. Using the specific California laboratory median control CV values, the probability of a MMEL violation based on the probability of a fail at or below 10 percent effect can be estimated for each laboratory. Table J-7 shows specific California laboratory median CV values and the probability of a MMEL violation (CVs were rounded up to match the incremental values shown in Table J-5). When using 10 replicates, all laboratories except Commercial Lab A are below a ½ percent chance of a MMEL violation based on the probability of fail at a 10 percent or less effect (Commercial Lab A has a seven percent chance). When increasing the number of replicates to 20, all of the laboratories are well below a ½ percent chance of a MMEL violation based on the probability of fail at a 10 percent or less effect.

Table J-7. Probability of a MMEL Violation Based on the False Positive Rate for Specific California Laboratories

Laboratory[1]	Median Control CV	p= Probability[2] of a Fail, PE ≤ 10% (N=10)	Probability of a MMEL Violation (N=10)	p= Probability[2] of a Fail, PE≤ 10% (N=20)	Probability of a MMEL Violation (N=20)
A-Commercial Laboratory[3]	0.23	0.2000	0.0720	0.0340	0.002273
B-Commercial Laboratory[3]	0.15	0.0480	0.0049	0.0030	0.000018
E-Commercial Laboratory[3]	0.11	0.0480	0.0049	0.0030	0.000018
F-San Jose Santa Clara Laboratory[3]	0.11	0.0480	0.0049	0.0030	0.000018
Commercial Laboratory #1	0.08	0.0020	0.0000	0.0000	0.000000
Commercial Laboratory #2	0.12	0.0480	0.0049	0.0030	0.000018
Commercial Laboratory #3	0.16	0.0480	0.0049	0.0030	0.000018
LACSD San Jose Creek Laboratory 2017-18	0.13	0.0480	0.0049	0.0030	0.000018

[1] In this analysis there are three unknown California laboratories (laboratories A, B, C) from the Fox et al. 2019 study and three unknown California laboratories (laboratories 1, 2, 3) analyzed by State Water Board staff. Because the names are not known, there is the possibility that the two sets include the same laboratories. Therefore, there could be as few as five separate laboratories total, and as many as eight included in this analysis. The data sets do not overlap temporally, therefore the analysis of the eight data sets are at a minimum, unique.

[2] Probabilities from or estimated from Table J-6 above.

[3] Laboratory data from Fox et al. 2019.

Probability of a Toxicity Reduction Evaluation Based on the False Positive Rate

Dr. Fox calculated the probability of two successive MMEL failures over a five-year period of time, based on the probabilities calculated in the previous two steps.

Table J-8 shows the range of probabilities. Continuing with the illustration for the median control CV of 0.15, using 10 replicates, there is a 4.8 percent probability of the TST declaring a fail when the percent effect is 10 percent or less, and a 0.49 percent probability of a MMEL violation based on fails at or below 10 percent effect. Based on these assumptions, there is a 0.00238 percent probability that a TRE will be required based on fails at or below 10 percent effect.

State Water Board staff used the probabilities provided by Dr. Fox in his memorandum and applied those probabilities to California laboratory CV data presented in this appendix. Table J-9 shows the results of this analysis. For seven of the eight laboratories, when using 10 replicates, there is less than 3 thousandths of one percent (0.00238% or 0.0000238) probability that a TRE would be

required due to fails below at or 10 percent effect when using 10 replicates. For one of the laboratories, Laboratory A, there is a still less than 9 thousandths of one percent (0.00894% or 0.0000894) probability that a TRE would be required because of high within-test variability (as expressed by a high median control CV of 0.23, which is shown in Table J-7).

Adding replicates to the toxicity test reduces the probability that a TRE would be required by TST fails at or below 10 percent effect. By using 20 replicates at Laboratory A, the probability of a single fail would meet the acceptable probability of a fail below at or 10 percent effect of five percent or less (Table J-5), and the TRE probability would be reduced from 0.00894 percent to 0.0003143529 percent.

Table J-8 Probabilities that a TRE would be Required Based on the False Positive Rate and a Replicate Number of 10

Probability of at least one run of 2 or more MMEL failures		
Probability *p* of a fail of a routine monitoring test	P_V[1], probability of MMEL violation in a calendar month	Probability of one or more runs (2 successive MMEL failures) in 60 months, resulting in TRE (based on P_V)[2]
0.02	0.000792	0.000000627264
0.03	0.001773	0.000003143529
0.04	0.003136	0.00000983
0.05	0.004875	0.0000238
0.06	0.006984	0.0000488
0.07	0.009457	0.0000894
0.08	0.012288	0.00015099
0.09	0.015471	0.00023935
0.1	0.019	0.000361
0.11	0.022869	0.00052299
0.12	0.027072	0.00073289
0.13	0.031603	0.00099875
0.14	0.036456	0.00132904
0.15	0.041625	0.00173264
0.16	0.047104	0.00221879
0.17	0.052887	0.00279703
0.18	0.058968	0.00347723
0.19	0.065341	0.00426945
0.2	0.072	0.005184
0.21	0.078939	0.00623137
0.22	0.086152	0.00742217
0.23	0.093633	0.00876714
0.24	0.101376	0.01027709
0.25	0.109375	0.01196289

[1] $P_V = p*p + p*(1-p)*p$

[2] Average of six simulation runs, each consisting of 10,000 simulated sequences of 60 months

Table J-9. California Laboratory Estimated Probabilities of a Two Successive MMEL Failures Over 60 Months Based on the Probability of a Fail Less than or Equal to 10 Percent Effect at N=10 and N=20

Laboratory[1]	N=10 Probability of MMEL violation each month	Probability of 2 successive MMEL failures in 60 months, resulting in TRE (N=10)	N=20 Probability of MMEL violation each month	Probability of 2 successive MMEL failures in 60 months, resulting in TRE (N=20)
A-Commercial Laboratory[2]	0.0720	0.0000894	0.0340	0.000003143529
B-Commercial Laboratory[2]	0.0049	0.0000238	0.0030	0.000003143529
E-Commercial Laboratory[2]	0.0049	0.0000238	0.0030	0.000003143529
F-San Jose Santa Clara Municipal Laboratory[2]	0.0049	0.0000238	0.0030	0.000003143529
Commercial Laboratory #1[3]	0.0000	0.0000	0.0000	0.0000
Commercial Laboratory #2	0.0049	0.0000238	0.0030	0.000003143529
Commercial Laboratory #3	0.0049	0.0000238	0.0030	0.000003143529
LACSD Municipal Laboratory 2017-18	0.0049	0.0000238	0.0030	0.000003143529

[1] In this analysis there are three unknown California laboratories (laboratories A, B, C) from the Fox et al. 2019 study and three unknown California laboratories (laboratories 1, 2, 3) analyzed by State Water Board staff. Because the names are not known, there is the possibility that the two sets include the same laboratories. Therefore, there could be as a few as five separate laboratories total, and as many as eight included in this analysis. The data sets do not overlap temporally, therefore the analysis of the eight data sets are at a minimum, unique.

[2] Laboratory data from Fox et al. 2019

[3] Probability of a fail at or below 10% effect was rounded up to 0.002 based on Table J-5. Fox memo presented the MMEL probability as 0.000, though the actual number is not zero, but extremely small.

References

Cuevas, Veronica. 2019. Los Angeles Regional Water Quality Control Board Personal Communication.

Diamond J, Denton D, Roberts J, Zheng L. 2013. Evaluation of the Test of Significant Toxicity for determining the toxicity of effluents and ambient water samples. Environmental Toxicology and Chemistry. 32(5): 1101–1108.

Fox J, Denton D, Diamond J, Stuber R. 2019. Comparison of False-Positive Rates of 2 Hypothesis-Test Approaches in Relation to Laboratory Toxicity Test Performance. Environmental Toxicology and Chemistry. 38(3):511-523.

Fox J. 2019. Memorandum: Probability of Failing TST and WET Maximum Monthly Effluent Limit. U.S. Environmental Protection Agency, Office of Research and Development, National Center for Environmental Assessment, Washington Division.

Hawaii State Department of Health. 2015. Hawaii's Nonpoint Source Management Plan 2015 to 2020.

Larry Walker Associates, Inc. 2018. *Ceriodaphnia dubia* Short-term Chronic Reproduction Test: Understanding the Probability of Incorrect Determinations of Toxicity in Non-toxic Samples.

Morris, Cristine. 2019. Los Angeles Regional Water Quality Control Board Personal Communication.

State Water Resources Control Board. 2011. Effluent, Stormwater, and Ambient Toxicity Test Drive Analysis of the Test of Significant Toxicity (TST). Sacramento, CA: Division of Water Quality.

State Water Resources Control Board. 2017. California Water Boards Quality Management Plan (Policy Guidance – Version 2.0). Sacramento, CA: Office of Information Management and Analysis.

Vazquez L. 2012. EPA's Test of Significant Toxicity: Impact on the Permit Compliance of Honolulu's Wastewater Treatment Plants. Water Quality Laboratory, Department of Environmental Services, City & County of Honolulu.

U.S. Court of Appeals, for the District of Columbia Circuit {Edison Elec. Inst., NACWA, et al. v. EPA, et al., No. 96-1062 (D.C. Cir. Dec. 10, 2004)}

U.S. Environmental Protection Agency. 2000. Understanding and Accounting for Method Variability in Whole Effluent Toxicity Applications Under the National Pollutant Discharge Elimination System Program. EPA 833-R-00-003. Washington, DC: Office of Wastewater Management.

U.S. Environmental Protection Agency. 2010. National Pollutant Discharge Elimination System Test of Significant Toxicity Technical Document. EPA 833/R-10-004. Washington, DC.