Ceriodaphnia dubia Quality Assurance Guidance Recommendations

Draft Report
August 31, 2023

Preface

This is a guidance manual with recommendations for improving the quality assurance of the *Ceriodaphnia dubia* survival and reproduction test. This document does not promulgate a new method or a formal change to an existing method. Instead, the recommendations are supported by a two and a half-year study analyzing data from each of the accredited toxicity testing laboratories for the State of California. This study was facilitated by the Southern California Coastal Water Research Project (SCCWRP) under contract to the State Water Resources Control Board (SWRCB) and the California Association of Sanitation Agencies (CASA). The project governance included a decision-making body, the Expert Science Panel, a five-member team of North American (non-Californian) experts familiar with performing *C. dubia* toxicity tests and representing the fields of aquatic toxicology and chemistry, statistics, and quality assurance programs related to environmental testing laboratories. The Expert Science Panel was assisted by a Stakeholder Advisory Committee representing 12 different sectors who utilize the *C. dubia* toxicity test for environmental management decision making, which includes a variety of regulated dischargers, state and federal regulators, and toxicity testing laboratories. All laboratory-specific results are anonymous as a condition of their participation in the study.

Members of the Expert Science Panel:

Robert Brent (James Madison University), Expert in Freshwater WET Testing

Howard Bailey (Nautilus Environmental), Expert in WET Testing

Teresa Norberg-King (Formerly US EPA), Expert in Freshwater WET Testing

Leana Van der Vliet (Environment and Climate Change Canada), Expert in quality assurance in single-species toxicity tests

A. John Bailer (Professor Emeritus, Miami University), Expert in Biostatistics

Members of the Stakeholder Advisory Committee:

Katie Fong, Representative for State Water Resources Control Board

Steven Boggs, Representative for California Environmental Laboratory Accreditation Program

Amelia Whitson (EPA Region IX), Representative for Federal Government

Rochelle Cameron (EPA Region IX), Alternate representative for Federal Government

Veronica Cuevas (Los Angeles Regional Water Quality Control Board), Representative for NPDES permits

Mitch Mysliwiec (LWA), Representative for California Associations of Sanitation Agencies (CASA)

Paul Bedore (Robertson-Bryan, Inc.), Alternate representative for CASA

Jian Peng (Orange County Public Works), Representative for California Stormwater Quality Association

Sarah Lopez (Central Coast Water Quality Preservation, Inc.), Representative for agricultural associations

Peter Arth (Enthalpy), Representative for private toxicity testing laboratories

Josh Westfall (Sanitation Districts of Los Angeles), Representative for public toxicity testing laboratories.

Annelisa Moe (Heal the Bay), Representative for Non-Governmental Organizations

SCCWRP Project Team

Alvina Mehinto and Ken Schiff- Toxicologists and Co-Principal Investigators responsible for facilitation of the two Committees

Darrin Greenstein – Quality Assurance Officer

David Gillett- Biostatistician

1. Executive Summary

The State of California (State) recently promulgated Toxicity Provisions for regulated dischargers (SWRCB 2020), part of which requires testing for the *Ceriodaphnia dubia* (*C. dubia*) survival and reproduction toxicity test (Method 1002.0, EPA 2002a,b,c), a test with decades of use after EPA ratified the approval of several test procedures for measuring the toxicity of effluents and receiving waters. More recently, stakeholders expressed concerns about variability associated with the *C. dubia* test method during the Toxicity Provision adoption process. While variability is an inherent property of test systems, high levels of variability could impair the usefulness of the *C. dubia* survival and reproduction test data for monitoring water quality and assessing compliance with discharge limits. This document summarizes the results of a two-year investigation into the level and sources of variability associated with the *C. dubia* survival and reproduction aquatic toxicity test. It also includes a series of recommendations to the state regulatory agency, the regulated parties, and the toxicity testing laboratories for maintaining and, in some cases, improving data quality for the *C. dubia* survival and reproduction toxicity test.

There are many potential sources of variability impacting toxicity testing results. For the *C. dubia* survival and reproduction toxicity test, these may include inherent differences in the response of individual test organisms and testing conditions that can alter organism condition and toxicant uptake and bioavailability (e.g., temperature, water chemistry and diet), health of the brood stock from which neonates are obtained, stress related to the presence of microorganisms in cultures and test solutions, and the skill and experience of staff performing laboratory techniques. To investigate the sources of variability in *C. dubia* test results, the following activities were performed:

- Inventory of lab techniques used by accredited laboratories
- Compilation and statistical analysis of historical data
- Baseline laboratory intercalibration study
- Laboratory visits and roundtable workshop
- Second laboratory intercalibration study

Inventory of laboratory techniques used by accredited laboratories

At the start of this study, 18 laboratories were accredited by the State. Three of the accredited laboratories were public agencies, 12 were private laboratories, and three were academic laboratories.

The inventory of laboratory techniques was created by 1) reviewing each laboratory's Standard Operating Procedures (SOPs) and Quality Assurance Plans (QAPs), 2) electronic questionnaires sent to each laboratory, and 3) one-on-one phone interviews with key laboratory staff.

The conclusion from the inventory compilation is that no two laboratories perform the *C. dubia* test in exactly the same manner. For example, when evaluating how labs create their laboratory dilution water, five laboratories use diluted mineral Perrier® or Evian® water and seven use reconstituted moderately hard water (MHW) or hard water as described in the EPA method manual (EPA, 2002). The others use a modified version of the MHW recipe or a completely different recipe. Lack of similarity could be observed for other laboratory techniques such as food and feeding, test set up (e.g., test chamber material and volume, light intensity, and photoperiod, etc.), culturing and brood boards, and reference toxicants. The most challenging area for evaluating sources of variability was when data was missing due to incomplete

recording of specific laboratory techniques for culturing and testing or when testing systems failed and provided no results.

Compilation and analysis of historical data

The study conducted an analysis of historical test results obtained from 17 of the 18 accredited laboratories in the State. In total, laboratory control data was compiled from 551 environmental sample tests along with 452 reference toxicant tests from the last three to five years. Variables of interest were selected for evaluation based on their relationships to test acceptability criteria (TAC) and general performance metrics, including survival, mean number of young produced by the controls, relative measures of variability include the coefficient of variation (CV) for mean number of young produced in the controls, and the percent minimum significant difference (PMSD) were calculated. Toxicity potency endpoint metrics (i.e., inhibition concentration (IC) associated with the 25% or 50% reduction in reproductive output, IC25 and IC50) were also included as measures of the toxic response to reference toxicants. These data were then compared within and among different laboratories to evaluate the level of consistency associated with individual laboratories as well as the comparability of different metrics across laboratories.

The observed range of estimated IC50s for the same reference toxicants within and among laboratories suggested some inconsistencies in implementing standard methodologies. The IC50s exhibited a greater than two-fold range of values across laboratories for sodium chloride (NaCl) or copper chloride (CuCl). Moreover, the level of variability in these metrics for several laboratories also suggested that standard practices were not being applied in a consistent manner within a laboratory. With respect to descriptive metrics, laboratory-specific mean of control reproduction varied among laboratories from 18.7 to 37.5 young per female, and individual test control reproduction ranged from <5 to >50 young per female. Of note, CVs associated with young production also varied appreciably, as did another measure of relative test variability that reflects a statistically detectable difference between responses in two test concentrations, the PMSD.

Collectively, data from the historical analysis suggested a broad range of procedural and quality control practices among different laboratories accredited by the State to perform the *C. dubia* chronic toxicity test. However, in this context, it is important to note that data from a subset of laboratories consistently met the desired performance metrics, suggesting that the discharge community has access to qualified testing laboratories capable of reliably performing the test on a routine basis. Regardless, the range of variability suggested that “standard” practices are not being applied consistently across or within laboratories. This is of particular concern given the availability of relevant guidance documents, and a decades-old State test accreditation program. Under these conditions, one might reasonably expect that most, if not all, of the certified laboratories would exhibit performance metrics consistent with regulatory guidance.

Given the range of variation in test performance metrics, the potential linkages between test performance and specific laboratory practices were evaluated. To accomplish this, a questionnaire was sent to each laboratory to request specific information on testing and culturing practices that might affect organism condition and test results. In cases where the responses were not clear, the questionnaires were followed up by phone or video interviews for clarification. All data were then analyzed to identify potential variables that might help explain variability in the test metrics.

Ultimately, no single or combination of laboratory practices appeared related to variability in control neonate production, variation in control neonate production, or reference toxicant endpoints. Descriptive statistics, multiple linear regression techniques, nor multivariable classification techniques (e.g., random forest) did not identify or suggest plausible explanations for differences between laboratories. One possible reason is that unique combinations of laboratory techniques that change over time are responsible for intra-and inter-variability. It should also be noted that each laboratory had a one-of-a-kind profile of experimental practice limiting the power of the analysis. Finally, it is reasonable to assume that these laboratories' characteristics are highly related and that identifying a single driving factor, or a few factors was not likely to occur.

Baseline interlaboratory comparison study

Given that the evaluation of the historical data did not reveal specific laboratory practices that accounted for an appreciable portion of the variability observed among the testing laboratories, an empirical approach was taken whereby selected sources of variability were controlled and then test procedures were documented in greater detail and consistency across laboratories. This baseline interlaboratory comparison study prepared different types of dilution water and distributed them to 12 different accredited laboratories, in conjunction with detailed data sheets designed to capture more of the underlying information associated with organism culture and testing procedures. The study was comprised of three separate testing events and used NaCl as the reference toxicant. The laboratories were asked to utilize their ongoing lab techniques as used for the tests submitted during the historical data compilation.

The baseline intercalibration provided similar results as the historical data analysis. After compiling the empirical data from 178 unspiked samples and another 60 spiked concentration-response tests, differences were observed between laboratories and no single (or combination of) factor(s) appeared to be related to the variability observed among laboratory results. The observed range of variability suggested the presence of inconsistencies between laboratories and in some cases TAC (survival and reproduction) were not met. Laboratory-specific mean reproduction in controls ranged from 14.9 to 40.2 young per female among laboratories, and wide variability was also observed for the reference toxicant metric, PMSD. Review of all the reference toxicant data (including those that did not meet test acceptability criteria), showed that the laboratory-specific grand means of IC50 for NaCl ranged more than two-fold among laboratories.

Statistical analysis identified a range of factors that were not responsible for the differences between laboratories but did not identify the technique or range of techniques responsible for the variation observed. The only factor that appeared related to neonate production in unspiked samples was the age of the female used to generate neonates for test set up; analyses revealed that test organisms obtained from older females tended to produce fewer neonates. However, these test organisms still produced sufficient neonates to pass test acceptability criteria, and this variable did not appear to affect the CV associated with neonate production. Consequently, both the experts and stakeholders expressed caution in over-interpreting these results.

Laboratory visits and roundtable workshop

The lack of identifiable factors that would explain the variability observed within and among laboratories suggested that underlying factors might be diverse, laboratory-specific, and not readily captured by data typically collected. Consequently, two members of the Expert Science Panel visited a subset of the laboratories representing a range of variability in test performance metrics to identify potential contributing factors. This was followed by a workshop attended by the Panel and 12 California accredited laboratories who participated in the baseline intercalibration study. The two-day roundtable workshop covered 20 topics in four areas of testing, culturing, food and feeding, testing, and documentation and recordkeeping. The laboratories actively engaged with the Panel and both parties agreed on a list of 16 items to standardize and apply in a second laboratory intercalibration study.

Second interlaboratory comparison study

The findings from the laboratory visits and the recommendations from the roundtable workshop were incorporated into a second interlaboratory study to determine if controlling identified potential causal factors would result in a commensurate reduction in variability of test performance metrics. Ten laboratories participated in the second intercalibration (1 public and 9 private laboratories), 9 of them had participated in the baseline intercalibration study. Laboratories that could not participate in the second ILD cited staffing issues.

This study was designed identically to the baseline intercalibration (i.e., same number and types of samples) with the exception that participating laboratories followed a common set of lab techniques defined in the roundtable workshop. This included use of neonates produced by 6- to 10-day old females to start the test, renewal, or termination of test boards daily at 24 h within a 2-h window, independent quantification of food density by the testing laboratories, food holding times of ≤ 7 for Yeast-Cerophyll®-Trout Chow (YCT) and ≤ 21 days for green algae, documentation of split broods on bench sheets at the time of observation. Testing was preceded with a training session for all participating laboratories for these standardized laboratory techniques.

Results from the second intercalibration indicated that two laboratories did not meet test acceptability criteria in at least one of the three testing rounds. Laboratory control mean reproduction ranged from 25 to 45 young per female. This was an improvement from the baseline study where three laboratories did not meet test acceptability criteria in at least one of the three testing rounds. For most laboratories, the CV for mean neonate production in their controls remained similar or showed a small reduction compared to the baseline intercalibration. Three laboratories, however, exhibited a wider distribution of CVs in the second intercalibration study compared to the baseline study. Samples tested as dilution series showed good intra-laboratory agreement, but differences in potency estimates for NaCl were observed among laboratories. Water quality, brood board health, food and age of females showed no obvious correlation with test outcome. Overall, one laboratory showed improvements that may be tied to the standardization of select test methods. However, the inconsistent laboratories remained inconsistent and some of the low-quality data were due to poor organism health or technical issues. Laboratories that exceeded expectations produced high quality and comparable data in both ILS.

Main findings

This study resulted in several findings related to the current implementation of the *C. dubia* reproduction toxicity test method by California accredited laboratories and the magnitude of variability within and among laboratories.

- Variability of test performance metrics was relatively high within certain laboratories and across the range of accredited laboratories.
    - At least one of the tests from five of the 12 accredited laboratories participating in the intercalibrations failed to meet the required TAC for survival or reproduction.
    - At least four of the 12 accredited laboratories performed within the Panel's expected level of consistency, providing a pool of qualified laboratories capable of performing the test on a routine basis.
- In general, no single aspect of culture or test procedures was identified that accounted for a large portion of the variability observed.
    - However, statistical analysis suggested that use of older adults in the brood boards may be associated with poorer performance of test organisms.
- Based on four onsite laboratory assessments and the subsequent workshop, it appeared that variability was most likely a function of multiple sources that occurred on an intermittent basis.
    - Contributing factors may have included preparation of water and diet and associated storage protocols, undesirable laboratory practices (e.g., conducting testing and culturing operations in same area), and poor husbandry procedures resulting in contamination of tests and cultures by micro-organisms.
    - Variability within individual laboratories was likely associated with lack of consistency in the application of specific procedures, potentially reflecting a lack of training and sufficient oversight by more experienced staff.
- Overall, standardization of select laboratory techniques produced modest improvements for laboratories with inconsistent or low historical performance.
    - Results suggest that at least one laboratory benefited from greater standardization of laboratory techniques.
    - Laboratories with historically high-quality data maintained high performance.
- The level of variability within and among laboratories for the *C. dubia* survival and reproduction test was inconsistent with their accreditation status, suggesting that the existing accreditation program including the proficiency testing, are not sufficient to achieve a uniform standard of quality across laboratories that ensures awareness of proper procedures, as well as their implementation.

Recommendations

The recommended guidance resulting from this two-year study falls into one of three categories: (1) Laboratory best practices, (2) Accreditation, and (3) Training.

Laboratory Best Practices

These recommendations are based on a review of the standard operating procedures, phone interviews, site visits and the roundtable discussion among laboratories and the ESP. The recommendations are directed largely at the laboratories. The recommendations are divided into "Must do" described in the promulgated method and EPA guideline for freshwater WET testing, and the "Should do" proposed to minimize variation but not required in the EPA method. It is important to note that EPA has already provided definitions of must/shall and may/should for WET toxicity test methods: "Words of obligation (EPA, 2000 Method guidance). WET test methods often state the procedure without a must or should, and in those instances, it is considered a directive. When WET method manuals use discretionary terms such as "may" or "should" the manual provide flexibility so that the laboratory analyst can optimize successful test completion.

Must do recommended guidance include:

- **Terminate the test when 60% of surviving females in the controls have had three broods, within a 2-h window (i.e., + or - 1 h) of test initiation time.**
- **Independently quantify food concentrations in stock bottles and record amounts added to each test container.**
- **Use source water produced according to the requirements of EPA freshwater WET test methods.**
- **Use known parentage with young from one adult for each concentration and use stratified random or complete randomization of all test cups.**

Should do recommended guidance include:

- **Conduct a detailed quantitative assessment of brood board health prior to testing.**
- **Document split broods on bench sheets daily at the time of the observations.**
- **Renew test solutions daily within + or - 2 h (i.e., 4-h window) of test initiation time.**
- **Update laboratory documentation.**
- **Store reagents to prepare the dilution waters and the reference toxicant appropriately.**

Accreditation

The ESP made recommendations on expanding the goals and implementation of the accreditation process (including proficiency testing) to ensure interlaboratory comparability. Since laboratory accreditation is the responsibility of the State, these recommendations are largely directed at the State accreditation program. These recommendations may increase operating costs for the laboratories, which will likely lead to increased costs to their testing clients. The recommendations are:

- **Increase the number and/or frequency of proficiency testing samples per year, following the model used in this project's intercalibration study.**
- **Collect and evaluate additional data associated with proficiency testing.**
- **Optimize laboratory audits to ensure effective and consistent implementation of best practices.**

Training Curriculum

Communication through regular WET testing training could ensure that the permittee or permitting authority has the information that they need to make informed decisions. Roundtable discussions and

public meetings with stakeholders highlighted the need to provide training materials for an improved understanding of method requirements and data quality objectives. Training recommendations are directed at the State, the laboratories, and at the regulated parties responsible for toxicity testing as a compliance requirement:

- **Implement auditors' training program.**
- **Implement training program with defined performance goals for all personnel involved in performing or reviewing the *C. dubia* test.**
- **Provide guidance to regulated parties to evaluate WET toxicity test data and understand the results.**

Limitations

While this study has produced more information on *C. dubia* inter-laboratory variation in more than 15 years, there are still a number of limitations to the conclusions and recommendations provided. These limitations fall into five categories:

- Limitations associated with the number of laboratories and the timing of the testing, which may not exhibit all the sources of variation possible in the *C dubia* reproduction test.
- Limitations quantifying the individual variability for each of the nine standardized laboratory best practices, but quantifying the improvement in variability cumulatively across all best practices.
- Limitations quantifying intra- and inter-laboratory variability associated with testing *C. dubia* reproduction in dilution water of varying hardness.
- Limitations on the number of toxicants quantifying intra- and inter-laboratory variability for concentration response in the *C. dubia* reproduction test.
- Limitations imposed by the study timeline and due dates, which impeded the Science Panel's opportunity to refine laboratory performance metrics, including developing additional guidance to define consistently performing laboratories.

Page Left Blank Intentionally

Table of Contents

List of Abbreviations and Definitions

ANOVA – Analysis Of Variance

C. dubia – Ceriodaphnia dubia

CV – Coefficient of Variation

DMW – Diluted Mineral Water; EPA recipe for moderately hard water using Perrier® or Evian®

ELAP – State of California Environmental Laboratory Accreditation Program

HW – Hard water is the EPA-developed recipe for reconstituted hard water to use for toxicity testing

IC – Inhibition Concentration is the toxicant concentration that would cause a given percent reduction in a non-quantal biological measurement for the test population.

IC25 – 25% inhibitory concentration

IC50 – 50% inhibitory concentration

ILS – Intercalibration Study refers to split-sample exercises among California accredited laboratories.

Inter-laboratory variability – the variability between laboratories, measured by comparing results from different laboratories using the same test method and the same test material.

Intra-laboratory variability – the variability within a laboratory, measured when tests are conducted using specific methods under constant conditions in the same laboratory. This variability includes within-test variability.

LC – Lethal Concentration is the toxicant concentration that would cause death in a given percent of the test population.

LC50 – 50% lethal concentration

LOEC – Lowest Observed Effect Concentration

MHW – Moderately Hard water is the EPA-developed recipe for reconstituted moderately hard water to use for toxicity testing.

NOEC – No Observed Effect Concentration

PMSD – Percent Minimum Significant Difference is the smallest significant difference from the control expressed as a percentage of the control mean.

PT – Proficiency Testing is a study conducted annually to assess laboratory performance in comparison to the other laboratories across the nation.

QAP – Quality Assurance Plan

SCCWRP – Southern California Coastal Water Research Project

SD – Standard Deviation

SOP – Standard Operating Procedures

SWRCB – State Water Resources Control Board

TAC – Test Acceptability Criteria

TNI- The NELAC Institute, a non-profit organization that operates the National Environmental Laboratory Accreditation Program (NELAP) and the National Environmental Proficiency Testing Program (NEPTP)

TST – Test of Significant Toxicity

U.S. EPA – United States Environmental Protection Agency

YCT – Yeast-Cerophyll-Trout chow

WET – Whole Effluent Toxicity

2. Introduction

The California State Water Board recently adopted Toxicity Provisions, which include numeric effluent limitations to protect California's enclosed bays, estuaries, and inland water bodies from contaminated discharges. The Toxicity Provisions also include a requirement to use the Test of Significant Toxicity (TST) statistical approach, which controls for both false positive and false negative error rates (Denton et al. 2011). This approach also "restated" the null and alternative hypotheses compared to traditional hypothesis tests; the null hypothesis of the TST being that the sample is toxic. Because of the controls on error rates and the restating of the null hypothesis, the TST approach is more likely to find a sample to be toxic if within-test variability is high. In using this approach, the State Water Resources Control Board (SWRCB) aims to incentivize dischargers to generate high-quality data (i.e., data with low within-test variability). However, dischargers have expressed concerns about the inherent variability in some of the Whole Effluent Toxicity (WET) tests included in the Toxicity Provisions such as the WET test for *Ceriodaphnia dubia* (*C. dubia*) survival and reproduction.

The *C. dubia* survival and reproduction test is a well-established and validated method and was first promulgated in October 1995 and finalized in 2002, nearly 20 years ago (U.S. EPA 2002a, b, c; U.S. EPA 2016). While the State Water Board has full confidence in the use of *C. dubia* for regulatory programs, they recognized that some laboratories may need to improve their implementation of the *C. dubia* method. For this reason, implementation of the monthly median effluent toxicity limitation for the *C. dubia* test has been delayed until January 1, 2024, for some dischargers as specified in the Toxicity Provisions. During this time, the State Water Board has committed to a study, in collaboration with stakeholders and laboratories, to evaluate laboratory performance, investigate factors that can lead to test variability and decrease confidence in assessments of toxicity or non-toxicity, and provide additional laboratory technique guidance to improve laboratory performance.

The WET test methods (EPA 2002a, b, c) allow laboratories some flexibility when implementing certain laboratory techniques. For example, for the *C. dubia* test method, different types of dilution waters (one made with salts, and one made with diluting a commercial mineral water) can be used. The appendix in the acute manual (EPA 2002a) describes the procedures for culturing and obtaining test organisms of *C. dubia.* In some instances, the promulgated method provides directions for what is to be done and also includes non-prescriptive test techniques, leaving laboratories to use their best professional judgement. In 2021, EPA Region 9 and California SWRCB held a *C. dubia* Lab Workshop to help California testing laboratories review and discuss the procedures. For the workshop, EPA Region 9 and the Office of Research and Development experts developed and provided the updated training in September 2021 prior to the study beginning to address test procedures.

The State of California Environmental Laboratory Accreditation Program (ELAP) accredits all laboratories conducting analysis for regulatory compliance purposes, including the *C. dubia* test. At the start of this study, there were 18 laboratories ELAP accredited to conduct the *C. dubia* test for California and 17 of them participated in the study (**Table 1-1**). Accreditation is based on the demonstration that laboratories are following the testing protocols, properly training their staff, keeping accurate records, demonstrating they can meet data quality objectives for internal reference toxicant samples and nationally distributed proficiency test (PT) samples. While the ELAP process demonstrates that a laboratory capably performs a test, it does not address test variability between laboratories or differences in lab techniques that are allowed by the protocols.

It has been hypothesized that the small methodological differences between laboratories may lead to intra- or inter-laboratory variability, which could influence test results. Previous studies have assessed the variability of the *C. dubia* test results within and among laboratories. In the early 2000s, an interlaboratory comparison exercise performed by the EPA found that 22 out of 122 *C. dubia* chronic tests did not meet TAC for survival or reproduction (EPA 2001a, b). The invalid tests were confined to 10 out of 34 participating laboratories. The study reported intra- and inter-laboratory coefficients of variation (CVs) for the IC25 values of effluent and receiving water split samples at 17% and 28%, respectively for the reproduction endpoint. More recently, a smaller interlaboratory comparison exercise was conducted in California to evaluate the reliability of *C. dubia* chronic test for stormwater toxicity (Schiff and Greenstein 2016). Of the nine laboratories that tested split samples of dilution water, three were considered "low comparability" based on three factors including test acceptability, intra-laboratory precision, and inter-laboratory precision. Lack of comparability among a minority of laboratories testing split samples of dilution water was also identified by others (Moore et al. 2000; Diamond et al. 2008). In Fox et al. (2019), NPDES data was examined from routine *C. dubia* survival and reproduction testing data from 2012 to 2015 that had been generated by eight California-accredited laboratories with tests being conducted using moderately hard, hard, and very hard water. The study compared two statistical approaches to determine the influence of laboratory test performance on the false-positive error rate. The study showed the need for laboratories to track their control CV and adopt measures to decrease within test variability (without enumerating how) and found that one laboratory that modified laboratory practices after 2012 showed their CV decreases from 0.31 to 0.17 (at the 75th percentile).

Various studies have focused on the causes of *C. dubia* test variability or ways to optimize the test. The main thrust of these studies has been the water used and organism feeding. Elphick et al. (2011) found that water hardness influenced the sensitivity of the organisms to chloride, with a decrease in toxicity observed as hardness increased. Other studies found acute toxicity associated with major ions (Na^+, K^+, Ca^{2+}, Mg2+, Cl^-, SO_4^{2-}, and HCO_3^-/CO_3^{2-}); salts can be a confounding factor both in natural waters and anthropogenically influenced waters (Mount et al. 2016, Erickson et al. 2017). Additionally, Mount et al. (2016) found that natural waters with low major ion concentrations caused *C. dubia* to be more sensitive to solutions of some salts. Chronic toxicity tests with *C. dubia* of major ion salts (Mount et al. 2019) showed a similar result to the acute study. Tests were conducted with 6 replicates and 9 closely spaced treatment concentrations were used to better support regression analysis of fairly steep response curves for modelling the data. For both the acute and chronic tests, the variability seemed to increase below an equivalent hardness of about 10-15 mg/L as $CaCO_3$.

One California laboratory conducted multiple studies to reduce sources of variability in their own *C. dubia* tests. The laboratory tested multiple dilution water types and sources and found that synthetic versus natural water had the most impact on reproductive variability, but it was small compared to feeding related aspects (Briden et al. 2017). In a study on the effects of water hardness, the California laboratory also found it had no impact on long-term culture performance (Clark and Briden 2018). However, the lab found that organism source and control/dilution water hardness might have an impact on test results. In two of six samples where both a soft and moderately hard water control was used, the interpretation of toxicity differed depending on which control the sample was compared to. The laboratory also conducted two studies looking at the effects of food quality. In the first study, the lab found that quality of the food had an impact on the test performance even if inferior quality food was only fed to culture animals, but higher quality was used during the test period (Jorgenson et al. 2017). The study also found that the quality

of the algae was the most important factor influencing control variability and was greater than control/culture water parameters, feeding density, food component, culture line, or analyst training. Source of the YCT did not appear to affect test control precision. In their second food study, they found that vendor sourced food was not necessarily of consistent quality (Prosser et al. 2018). The laboratory concluded that it was important to run quality control (QC) tests before using the food in cultures or tests. Little difference was found in reproduction based on variable food components, but when larger volumes of trout chow were digested a negative impact on test performance was observed. The laboratory also noted that the EPA recommendation of a 2-week shelf life for a *Selenastrum* batch may be too restrictive. Visual and olfactory observation of each batch were important to determine shelf life.

Table 2-1. Toxicity testing laboratories accredited in the state of California at the start of the study in 2021.

Lab Name	**Lab Type**
ELAP accredited laboratories in California	
49er Water Laboratory	Private
Aqua-Science	Private
Aquatic Bioassay & Consulting Laboratories, Inc.	Private
Aquatic Testing Laboratories	Private
Aquatic Toxicology Laboratory, Aquatic Health Program, UC Davis	Academic
Enthalpy Analytical, LLC	Private
Environmental Monitoring Division (EMD) at Hyperion Treatment Plant	Public
Inland Empire Utilities Agency Laboratory	Public
MBC Aquatic Sciences	Private
McCampbell Analytical, Inc.	Private
Pacific EcoRisk	Private
San Jose Creek Water Quality Laboratory	Public
Wood Environment & Infrastructure Solutions, Inc.	Private
ELAP accredited laboratories outside of California	
EcoAnalysts, Inc.	Private
Eurofins TestAmerica - Corvallis (ASL)	Private
GEI Consultants, Inc.	Private
Tetra Tech's Ecological Testing Facility	Private

3. Study Objectives

The objective of this study was to build on previous efforts and investigate all possible sources of variability in the *C. dubia* reproduction test conducted by California-accredited laboratories. The goal was to provide laboratory technique guidance to: (a) improve the consistency of the execution of the *C. dubia* test method to achieve improved precision within each testing laboratory; and (b) improve the consistency and comparability of *C. dubia* test results among testing laboratories, while retaining the necessary flexibility for environmental relevance.

The study aimed to answer the following questions:

1. What are the *C. dubia* chronic survival and reproduction toxicity test laboratory techniques used by Environmental Laboratory Accreditation Program (ELAP) accredited laboratories in the state of California?
2. How does variability in control reproduction and/or reference toxicant response in the *C. dubia* chronic survival and reproduction toxicity test compare amongst intra- and inter-laboratory technique differences used by ELAP accredited laboratories?
3. Does standardizing differences in the *C. dubia* chronic toxicity test laboratory techniques reduce intra- and inter-laboratory variability in control reproduction and/or reference toxicant response?

Based on the results of this study, a list of suggested best practices for the *C. dubia* reproduction test laboratory techniques were developed.

Note that this study was not designed to address or quantify false negative or false positive rates for detecting toxicity from known or unknown samples. It was also not expected to eliminate all variability from the test method. Finally, it should be noted that this study was not designed to address aspects of testing that may be more effectively dealt with by appropriate study design: e.g., ion, hardness, or conductivity controls in cases where those variables have the potential to affect test outcomes, but do not represent environmental risks.

4. General Approach

Six tasks were used to address the study objectives. These tasks were sequential with each one informing the details of the next.

1. Create a governance structure to oversee the study
2. Analyze historical data and existing lab techniques to identify sources of variability
3. Conduct a baseline intercalibration study to quantify variability within and among laboratories
4. Agree on a standardized list of laboratory procedures
5. Evaluate the efficacy of standardized laboratory techniques in reducing intra- and inter-variability via a second intercalibration study
6. Provide final recommended guidance in a Final Report

The first Task created a two-tiered governance structure to ensure transparency and technical rigor. One tier was a Stakeholder Committee comprised of representatives from sectors potentially impacted by the study results. The second tier was an independent Expert Science Panel comprised of scientists experienced in the *C. dubia* test method, biostatistics, and data quality measures, and with no potential

conflict with study results. The Expert Science Panel was the final decision-making body. The Stakeholder Committee provided valuable input and context for recommended guidance implementation.

The second Task was comprised of two subtasks. The first subtask was compiling an inventory of lab techniques used by ELAP accredited laboratories. The inventory elucidated the level of comparability and differences in test implementation. The second subtask focused on compiling historical testing data from the ELAP accredited laboratories to quantify the level of variability within and among laboratories. Approximately 1,000 tests were compiled from all but one of the ELAP accredited laboratories for this subtask. The inter- and intra-laboratory variability was assessed based on the reproductive endpoints of the test method (e.g., average number of neonates per female). The differences in lab techniques were compared to the lab test results to attempt relating which lab techniques might account for the observed variability in the test outcomes.

The third Task collected new data using a baseline intercalibration study using well-homogenized, split samples to assess intra- and inter-laboratory variability. The split sample analysis for this subtask supplements the historical data analyses and was intended to confirm possible sources of test variation. The basic study design for the baseline intercalibration was to remove the variability associated with samples (which was not possible in the historical data) and quantify the variability introduced by the individual laboratories. Laboratories utilized their existing protocols for all baseline intercalibration testing.

The fourth Task focused on identifying lab practices to standardize to reduce the interlaboratory variability. This task was accomplished using two sub-tasks. The first subtask consisted of on-site laboratory visits by a subset of the Expert Science Panel. These Panel members observed each laboratory's culturing, dilution water preparation, food and feeding, test implementation, quality assurance, amongst other activities. Four labs were visited that comprised a range of size and consistency of quality from the baseline intercalibration. The second subtask was a roundtable workshop convening of all the intercalibration participating laboratories. Based on the differences in lab procedures identified during the historical data inventory, the baseline intercalibration, and the lab visits, the goal of the roundtable workshop was to achieve consensus on what lab procedures to standardize for Task five.

The fifth Task conducted a second intercalibration, which mirrored the baseline intercalibration, except for the laboratories standardized the list of lab procedures agreed to during the roundtable workshop. These lab procedures were well-documented, and laboratories trained on how to implement them. The goal was to assess if the standardized lab procedures in Task four improved laboratory intra- and interlaboratory variability.

The sixth Task documents the study and lists the final recommended guidance on improving intra- and inter-laboratory variability. This report is the culmination of task six.

5. Evaluation of Intra-Laboratory and Inter-Laboratory Variability

Several types of information and datasets were evaluated during the two-year project to assess laboratory performance and determine the potential sources of variability.

5.1. Inventory of standard operating procedures, quality assurance plan and historical testing data from California's accredited laboratories

To investigate factors that can lead to test variability, an inventory of laboratory techniques and historical data was created focusing on culture and test conditions, and performance data for control samples and reference toxicant. **Table 5-1** presents the parameters collected. Out of the 18 accredited laboratories, one did not participate due to lack of *C. dubia* data available (i.e., less than 15 test tests over a 10-year period), and two laboratories provided incomplete information. To compile culture and test condition parameters, a review of the laboratory documentation such as Standard Operating Procedures (SOPs) and Quality Assurance Plans (QAP) was conducted. The performance data collected focused on the last 30 tests or up to 3 years for control samples associated with environmental test samples as well as reference toxicant concentration- response data. Control samples data were used to assess the lab's ability to perform the test, while reference toxicant data were used to assess reproducibility of test organism response. For each test, raw data from individual replicates was collected (i.e., daily neonate counts, survival) along with daily water quality data (i.e., hardness, alkalinity, pH, temperature) and other relevant metadata (e.g., brood board health). Environmental sample toxic response data were not used because there were no expectations of performance and data could not be compared among laboratories. It should be noted that the compiled tests included all samples regardless of whether the tests met TAC or not. The extracted data were hand-entered in a custom database and two independent audits were performed to assess completeness, accuracy, and variability. To verify the information compiled and collect additional data, phone interviews were conducted with key personnel from each laboratory. A survey questionnaire was submitted to the laboratories prior to the phone call and used during the discussion.

Notable differences were observed in all key parts of the test method including dilution water, test termination trigger and feeding techniques among the laboratories (**Appendix A Tables A2 and A3**). The test termination trigger is an important laboratory technique described in the promulgated method (Section 13.10.9.1), and it is specified that test termination "must be completed when 60% of the females or more have produced three broods". While most laboratories followed the requirement, the method to determine when the reproduction threshold is met varied greatly. Some laboratories used a strict time window daily while others checked periodically throughout the day to determine if the 60% threshold had been reached. Other laboratories documented using a higher percentage of females having produced three broods to ensure sufficient neonate production (e.g., 70% or 80% of females having three broods). One laboratory implemented a 7-day test consistently independently of the reproduction threshold. Control charts were also different as laboratories used either sodium chloride (NaCl) or copper chloride (CuCl) in different concentrations to prepare their serial dilutions. One laboratory reported using zinc sulfate as the reference toxicant.

Table 5-1. Laboratory techniques and performance variables compiled from California accredited laboratories.

Laboratory practices	**Testing and Performance Variables Recorded**
Origin of brood stock	Age window at test initiation
Age of culture	Time to reproduction
Culture renewal frequency	Test termination trigger
Dilution water recipe	Test termination window
Source water	Test duration (days)
Dilution water shelf-time	Number of neonates per female per replicate
Reference toxicant name and source	Number of replicate test chambers
YCT vendor, shelf-time	Survival of control females per replicate
YCT concentration in culture and test chamber	Neonate production in control samples (mean, CV)
Algal species	Reference toxicant, LC50
Algae vendor or recipe, shelf time	Reference toxicant, IC50
Algae concentration in culture and test chamber	Percent minimum significant difference
Feeding frequency	Water hardness
Lab air temperature	Water conductivity
Photoperiod	Water dissolved oxygen
Light source	Water temperature
Lab air temperature	Water pH
Sample volume in test chamber	Water alkalinity
Test chamber material, volume, diameter	

The EPA manual allows some flexibility in source water and dilution water recipes and, as a result, most California accredited laboratories have reported using modified dilution water recipes. Eight laboratories appeared to use one of the dilution water recipes specified in the EPA manual, either DMW or MHW with or without selenium. However, further investigations into the preparation of source water showed that some of them may not be using high quality source water (with a resistance ≥ 18 megaohm-cm). Other laboratories who used the MHW recipe often added different amounts of vitamins and/or selenium or adjusted the salts ratio (referred to herein as modified EPA recipe). Only one laboratory used EPA hard water (HW) for their culture and laboratory controls because the hardness of their test samples is usually outside of the range of MHW targeted in the EPA manual. It should be noted that two of the accredited laboratories did not use any of the dilution water recipes described in the promulgated method (Labs I and K). Food source, preparation and distribution were also different among laboratories. The feeding regime was also vastly different, as laboratories used different vendors and laboratory techniques (purchased or in-house) to produce their YCT and green algae *Raphidocelis subcapitata* stocks and feed them to *C. dubia* cultures. Previous research showed that the quality of food can affect both the number of neonates produced and the variability between tests within a single laboratory (Jorgenson et al. 2017). Other notable differences included material and size of test chambers as well as sample volume used in the test chambers.

A total of 551 sets of control data and 452 reference toxicants tests (**Table 5-2**) were entered in the database. Note that a 'set' is comprised of the reproductive output of 10 or 20 replicate test cups. Only two laboratories did not provide a complete set of control data, Lab H and Lab J. These two laboratories did not participate in the subsequent tasks of this project. Test data collected were typically from 10 replicate chambers, except two laboratories that occasionally conducted their tests using 20 replicates for the submitted control data. All reference toxicant data consisted of five dilutions minimum tested in 10 replicates. Approximately half of the laboratories had ~30 or more tests available within a 1.5 to 3-year period. The other half of the laboratories reported conducting less frequently and submitted between 6 and 25 sets of control or reference toxicant test data. One lab, Lab B, provided 7 years of data. Raw data were used to calculate test endpoints such as mean neonate production per surviving female, mean survival of females in controls, IC25/50 for reference toxicant at test termination (**Appendix A**).

Table 5-2. Inventory of historical data compiled from 17 California accredited laboratories.

Labs	Total number of laboratory control tests	Number of laboratory control tests with 10 replicates	Number of laboratory control tests with 20 replicates	Number of reference toxicant tests
A	48	48	0	31
B	48	48	0	47
C	28	28	0	28
D	19	19	0	6
E	49	24	25	30
F	45	37	8	30
G	7	7	0	22
H*	0	0	0	17
I	30	30	0	30
J	7	7	0	21
K	19	19	0	15
L	27	27	0	30
M	59	59	0	34
N	30	30	0	30
O	30	30	0	30
P	80	1	79	28
Q	25	25	0	23
Total	**551**	**439**	**112**	**452**

* Lab H did not respond to request for laboratory control data.

5.2. Evaluating laboratory performance.

Evaluating test performance and consistency is an important component of any monitoring program to ensure that all testing laboratories have a similar level of competency. Initially, a statistical approach was applied (e.g., Analysis of Variance, ANOVA). However, the Panel noted that high variability within and among laboratories limited the ability to detect any significant changes. Therefore, the Panel used a combination of metrics on biological, variability/uncertainty and potency endpoints and the magnitude of change to assess laboratory performance. The specific criteria and metrics used are listed below.

- Biological metrics (i.e., TAC) provide information on control test organisms, and reflect culture health and good laboratory practices. Meeting TAC is a requirement of the method. Several factors, such as dilution water, feeding regime, and environmental conditions can affect TAC.
    - Survival ≥ 80 percent in laboratory controls.
    - Mean number of neonates per surviving female ≥ 15 in laboratory controls.

- Metrics of variability and uncertainty provide information on consistency of test results within a laboratory. These were evaluated using both laboratory controls and reference toxicant samples, where sample response expectations exist.
    - CV for mean number of neonates per female in controls ≤ 0.2, i.e., the standard deviation in the number of neonates is less than or equal to 20% of the mean number of neonates in the group. This value is consistent with observed variability described in EPA (2000a) and Fox et al. (2019).
    - PMSD for individual reference toxicant tests ≤25 which corresponds to the 50^{th} percentile using data from 30 laboratories across the nation (EPA 2000a).

- Toxicity metrics focused on potency estimates for the reference toxicant and comparability among laboratories. This metric can be evaluated in different ways depending on the distribution of the data and the desired confidence interval. The metric below is used in the present study as an example of how it can be applied to assess interlaboratory comparability.
    - IC50s within the 25^{th} and 75^{th} percentile of all laboratories. There is currently no guideline for this metric in the EPA manual. The only guidance available is to reevaluate IC50 data that fall outside of two standard deviations within a laboratory.

The use of performance metrics will benefit all parties involved to describe, monitor, and communicate clearly the acceptable level of variability for this test. Testing laboratories can use the data as indicators of test organism condition and implementation of good laboratory practices. Regulators and regulated dischargers can use the information to increase confidence in test results, facilitate data interpretation and improve compliance with water quality monitoring objectives. It is important to note that the metrics used in the analysis of the *C. dubia* datasets are not inclusive or intended to be used as definitive guidelines for laboratory assessment. The State accreditation program could further refine the acceptance metrics and potentially include additional criteria (e.g., IC25, ratio of IC25 and IC50, LOEC, NOEC) based on the goals and objectives of the accreditation program.

5.3. Analysis of historical data

Historical data was evaluated in two ways. Key biological, variability/uncertainty metrics and potency estimates were compared within and among laboratories to assess overall laboratory performance. The potential relationships between laboratory techniques, test factors (e.g., water quality) and test outcomes were investigated using a variety of linear and non-linear modelling approaches detailed below.

Laboratory performance

Test acceptability criteria for the *C. dubia* chronic toxicity test focus on the performance of laboratory controls and require ≥80% survival and mean production of 15 young per surviving female. Historical data analysis showed that all California-accredited laboratories were able to meet the TAC for survival consistently, with 15 out of 17 laboratories reporting ≥ 90% in over 90% of their tests. Neonate production, however, was more variable. While most laboratories did achieve the TAC for reproduction, a 2- to 3-fold difference in mean control reproduction was observed among tests within a laboratory (**Figure 5-1**). Out of the 17 laboratories, two (Lab H and J) had an average mean reproduction close to the TAC and did not meet the criterion in more than 1 in 10 (i.e., ~12%) of their tests. The others typically had means of ≥ 20 young per surviving female in over 9 in 10 (90%) of their tests.

Meeting TAC alone is not sufficient to assess control and overall laboratory performance. Thus, the Panel critically evaluated metrics of variability, starting with the CV. Mean CVs for individual tests ranged from 0.14 for Lab A to 0.32 for Labs B and G. Wide differences were also observed within several laboratories (**Figure 5-2**). Further analysis was performed using a CV ≤ 0.20 suggested by the Panel as a metric of good laboratory performance. This value implies that a laboratory can achieve a standard deviation in the average number of young less than or equal to 20% of the mean of this average. This is within the range of CVs observed in previous studies assessing the variability of *C. dubia* test results (EPA 2001b, Fox et al. 2019). Such a level would be appropriate for confirming the presence of toxicity in environmental and other test samples. Only two laboratories (Lab A and Lab P) met this criterion in 8 of 10 of their tests. It should be noted that Lab A was the only laboratory with a 75th percentile of CV ~ 0.15 indicating that this laboratory had one of the most consistent control reproduction datasets. The 75th percentile of CV for Lab P was ~ 0.2. Approximately 1/3 of laboratories (6 out of 17) were also able to achieve a CV of < 0.2 on occasion, with 50-70% of their tests meeting the suggested guideline. These data suggest that a "long-term" average CV of ≤ 0.2 is achievable. The remaining laboratories (9 out of 17) exhibited higher variability among tests for control reproduction suggesting that these laboratories could benefit from greater level of standardization and guidance to improve data consistency.

Another statistical criterion used in hypothesis testing to assess toxicity test performance is PMSD (Denton et al. 2003). The EPA method guidance (2000a) suggests that PMSD be monitored as part of a testing laboratory's ongoing QA program. This metric describes intra-test variability and represents the smallest percent difference that can be statistically detected when a test of mean differences between two concentration groups is conducted. While there is no target PMSD value for compliance, the EPA has suggested a PMSD of 37%, which was the 90th percentile calculated using data from a national intercalibration study using a reference toxicant approach (EPA 2000). Consistent with the EPA study, the 90th percentile for the data compiled in the current project was 36%. When comparing the mean PMSD per laboratory to the EPA "upper bound" of ≤ 37%, all laboratories were able to meet this threshold in at least 6 out of 10 of their tests (**Figure 5-3**). However, the actual PMSD values ranged between 6 and 159

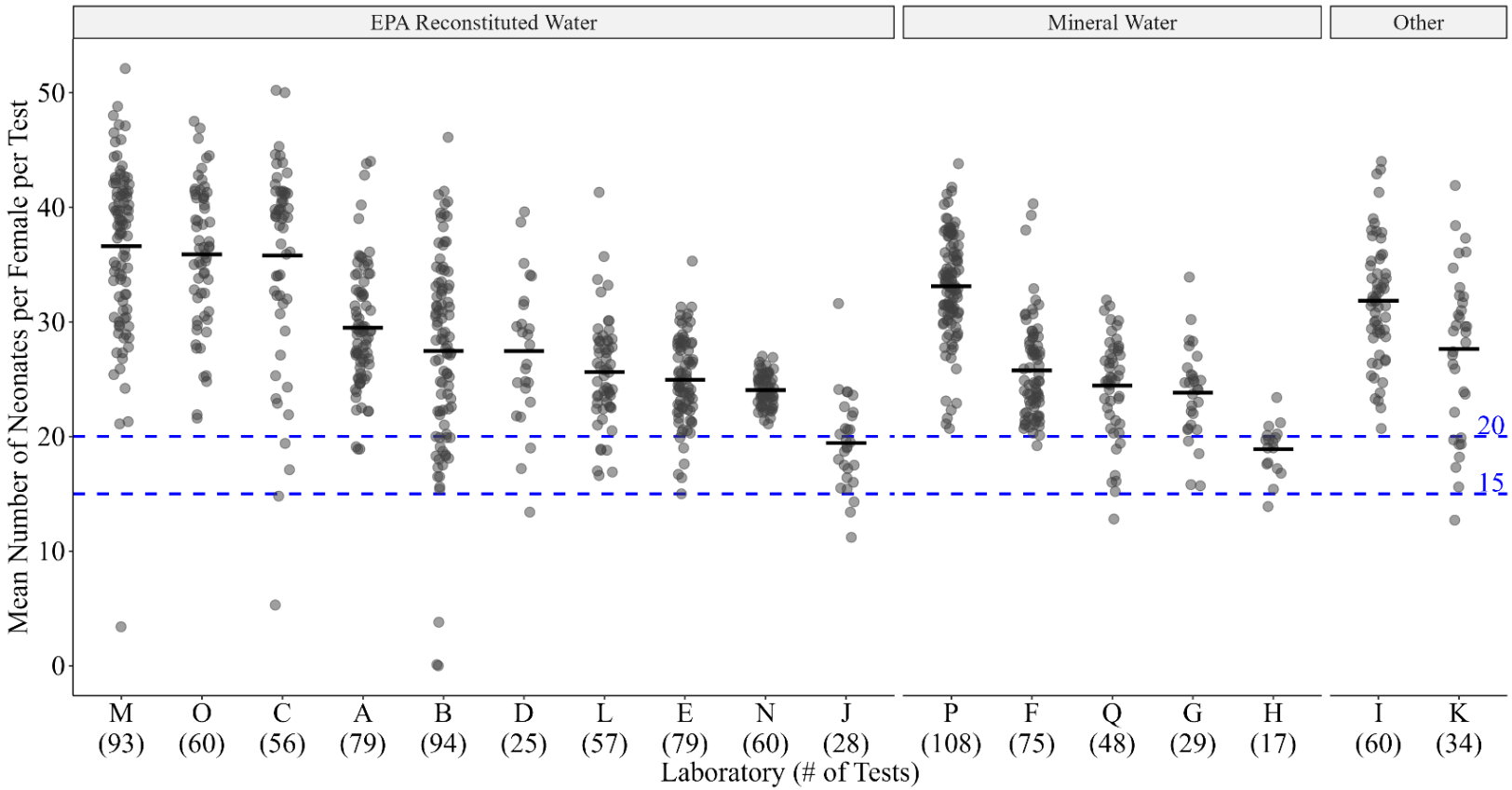
among laboratories. Six laboratories (B, C, H, I, J, K) had a noticeably wide range of PMSDs by a factor of 10 to 24, with CVs between 0.8 and 1.3 for these laboratories (**Appendix A Table A6)**. This suggests that a guideline of ≤ 37% may not be representative of the performance of a competent and consistent laboratory.

When setting the PMSD target value at ≤ 25%, which corresponds to the 50th percentile for the EPA study (EPA 2000), seven laboratories met the target value in over 75% of their reference toxicant tests and five additional laboratories met the target in 50 to 70% of their data. It should be noted that laboratories consistently meeting TAC and a CV of ≤ 0.2 for reproduction (Labs A, F, P) also had the greatest level of consistency in PMSD values. Overall, these data showed that ~2/3 of the laboratories were able to detect a 25% inhibition level relative to the control on a consistent basis, whereas the remaining laboratories tended to be characterized by high variability and would likely benefit from a rigorous review of culture conditions and testing procedures.

Most laboratories (10) used NaCl as their reference toxicant, six laboratories used a form of copper and one laboratory used zinc. Therefore, reference toxicant data could not be used to assess interlaboratory agreements among all California accredited laboratories. Mean IC50s ranged between 15 and 65 µg/L for copper and between 1000 and 2400 mg/L for NaCl (**Figure 5-4**). Most laboratories (8 out of 10) using NaCl had mean IC50 within the 25th and 75th percentiles of all reference toxicant tests. Two outliers were identified; the mean IC50 for Lab B fell within the 10th and 25th percentile and Lab K fell outside of the 90th percentile of all reference toxicant tests. This suggests a reasonable level of comparability among most laboratories using NaCl. The distribution of IC50s for copper was slightly more variable. Half of the laboratories had mean IC50s within the 25th and 75th percentile of all copper reference toxicity tests, two laboratories were between 25th and 10th percentiles, and one laboratory fell within the 75th and 90th percentiles.

Water quality parameters, including temperature, dissolved oxygen, pH, conductivity, hardness, and alkalinity were qualitatively examined (**Appendix A Table A5**). Mean measurements were comparable among laboratories, except Lab M that had a lower mean dissolved oxygen and Lab F that had lower mean alkalinity. Lab B and M had wide ranges of dissolved oxygen and temperature, respectively. Lab E, H and K also had highly variable ranges of hardness, conductivity and/or alkalinity. Lab C water quality data was not compared to the other laboratories because they routinely use HW which has higher hardness, conductivity, and alkalinity.

Figure 5-1. Mean number of neonates per female for each submitted test from the historical dataset. Data is organized by the type of dilution water used by the laboratory for their test controls. Laboratories are in order of high to low mean values within each water type.



Note that most laboratories using the EPA reconstituted water add selenium and/or vitamins (see Appendix A, Table A2).

Figure 5-2. Coefficient of variation for the mean number of neonates per female for each submitted test from the historical dataset. Data is organized by the type of dilution water the laboratory uses in their controls. Laboratories are ordered from high to low mean values within each water type.

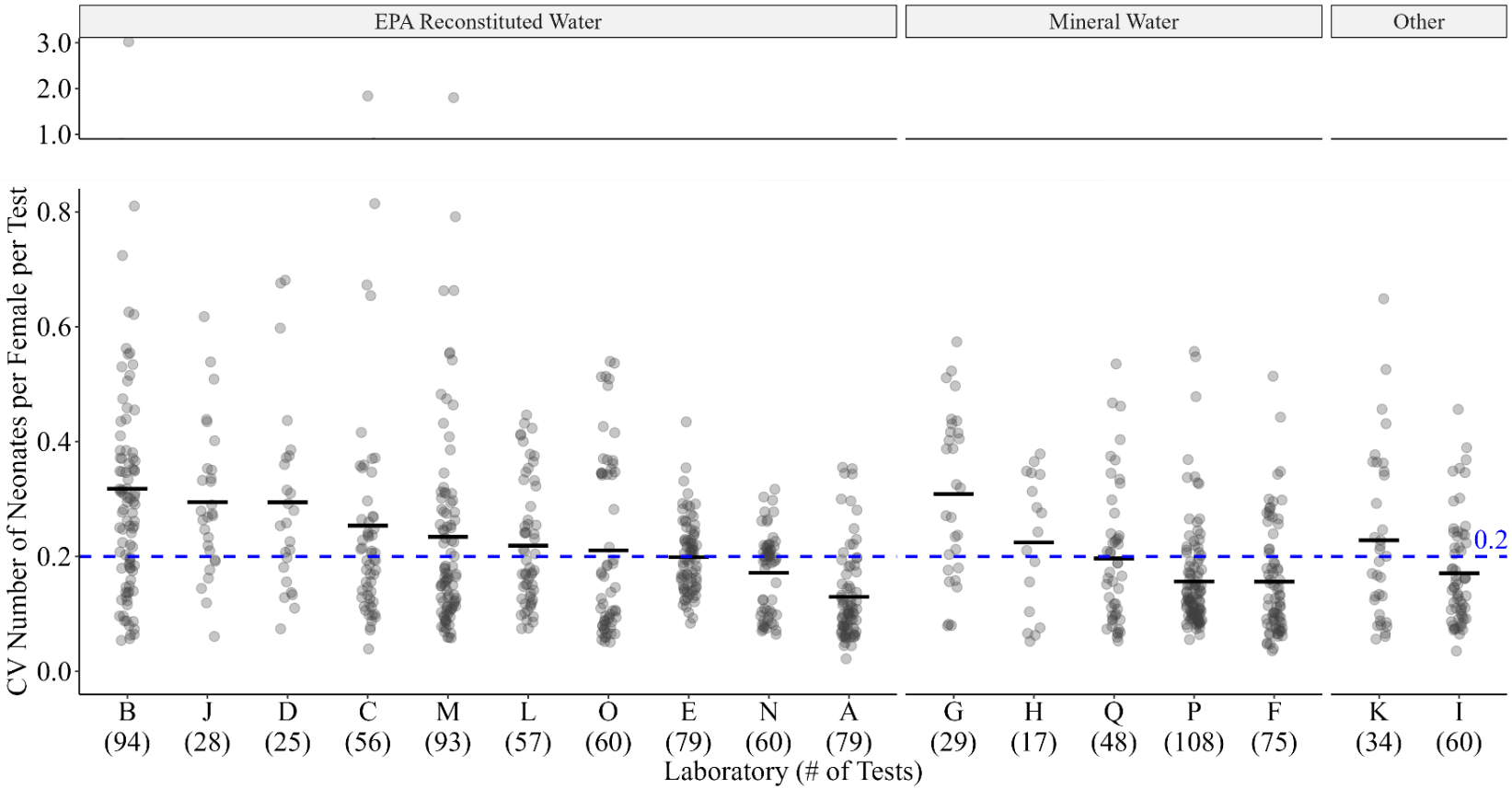


Figure 5-3. Percent minimum significant difference (PMSD) values from the reference toxicant data in the historical dataset. Reference lines indicate proposed laboratory performance criteria. Line at 37 is based on 90th percentile of data from an EPA (2000) study. Line at 25 is based on the Science Panel's suggested guideline and is the 50th percentile from EPA (2000). Laboratories are ordered based on their mean IC50 for each reference toxicant type.

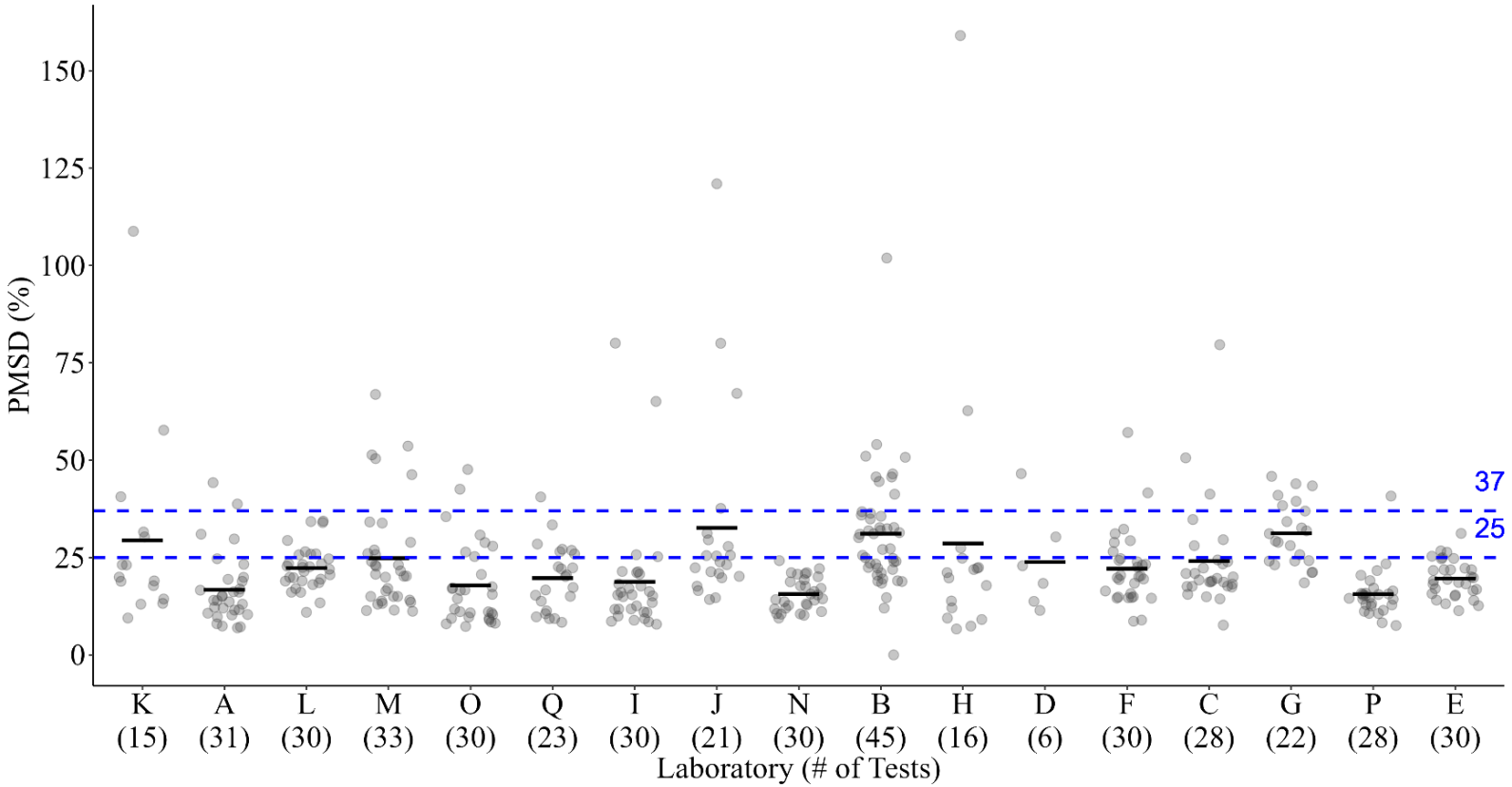
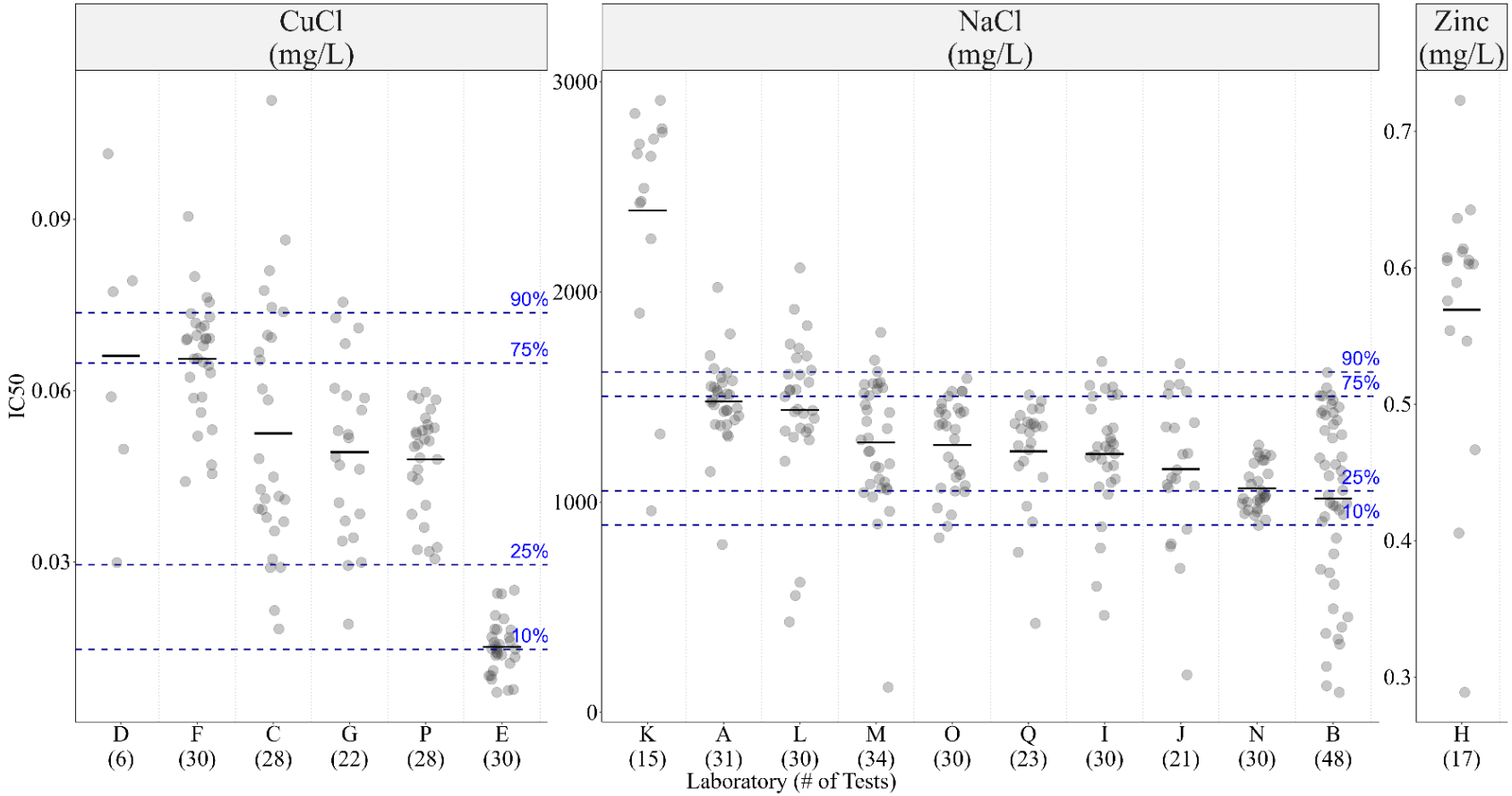


Figure 5-4. IC50 data from the historical dataset. The reference lines are percentiles of the data from each reference toxicant type. Larger, colored dots represent the mean value for each laboratory. The laboratories are ordered by highest to lowest IC50.



Investigating sources of variability in test outcomes.

The relative influence of laboratory techniques and test conditions (listed in **Table 5-1**) on inter- and intra-laboratory patterns of the means and CVs of control neonate production per test were investigated using univariate and multivariate linear modelling. The results indicated that there were no strong relationships between any one of the different lab techniques / test conditions with either mean or CV of control neonate production across the different laboratories.

To further explore important factors in neonate production across the historical data, random forest regression models (consensus-based, non-linear modelling) (Breiman 2001; Biau and Scornet 2017) of the mean and CV of control neonate production as a function of all the lab techniques / test conditions concurrently and interactively were created. Specifically, regression models were created for each laboratory with either mean or CV of neonate production as the response variable and all metrics in **Table 5-1** used as predictor variables. Variable importance measures extracted from each laboratory's random forest model were ranked by importance and compiled. Measures of test water quality (e.g., temperature, pH) and composition (e.g., hardness, alkalinity) were highly ranked for a number of laboratories, suggesting that they could potentially be important drivers of both the mean and CV of control neonate production. However, there was no compelling agreement in the models among all laboratories. The variable importance results also suggested that the age of the females used in the could be an important factor in mean control neonate production.

The results of the random forest models were used to build more structured, and potentially more diagnostic, multivariate general linear models of water quality, water composition, and female age on mean and CV of neonate production. However, due to the incomplete data record and high degree of heterogeneity in test water and female brood stock within and between labs, no conclusive results could be identified as to the sources of the observed variability in control neonate production among the labs in their historical test data.

Despite the lack of relationships between lab techniques / test conditions and mean or CV of test neonate production, the results did allow the Panel and the Stakeholder Committee to come to a consensus that controlling for water composition, as well as better reporting of test condition data (brood board characteristics, feeding regimes, and testing conditions) could produce a more complete and less "noisy" data set on which more diagnostic analyses could be conducted.

5.4. Split-sample intercalibration exercises among California's accredited laboratories

A three-step approach was used to assess the effects of standardizing laboratory practices on laboratory consistency and comparability. First, the laboratories participated in an interlaboratory study (referred to hereafter as baseline ILS) and tested split samples using their own protocols and provided more detailed data that may not be routinely collected and reported. **Appendix B** includes an excerpt of the QAP developed for this exercise and provides more information on the study design. Based on the results of the baseline ILS, the Panel conducted site visits in four of the laboratories and developed a list of topics to be discussed with the participating laboratories. A roundtable workshop was hosted, where laboratories shared their practices with the Panel and the group agreed on a set of practices to standardize in a second intercalibration study (referred to hereafter as second ILS). In the second ILS, laboratories tested split

samples following the standardized *C. dubia* toxicity testing techniques agreed upon in the Workshop. **Appendix C** provides an overview of the key study elements for the second ILS testing.

Twelve California-accredited laboratories (2 public and 10 private) participated in the baseline ILS and nine (1 public and 8 private) in the second ILS. Those who could not participate in one of the two ILS cited issues with their cultures or lack of staff and time. These are identified in the analyses performed. The baseline and second ILS followed the same testing design, with three rounds of testing per exercise. For each round, SCCWRP prepared and shipped four types of split samples to the laboratories.

- Sample 1: MHW water recipe tested at full strength (i.e., 100%, no serial dilution needed). This sample was tested along with one laboratory control consisting of their own dilution water recipe.
- Sample 2A: DMW with Perrier® water tested at full strength (i.e., 100%). This sample was tested along with one laboratory control consisting of their own dilution water recipe.
- Sample 2B-F: Five concentrations of NaCl diluted in DMW with Perrier®. The five samples were prepared at SCCWRP according to the procedure described in the study QAP. These samples were tested as is with no additional sample dilution allowed, along with one laboratory control consisting of their own dilution water recipe.
- Sample 3: NaCl was provided as a solid to each laboratory with detailed instructions to prepare five dilutions using their own dilution water. The laboratory-prepared serial dilution was tested along with one laboratory control consisting of their own dilution water recipe.

It should be noted that all samples were tested in 10 replicate test chambers and the tests were all conducted for 8 days with neonate counts recorded daily.

Data were submitted electronically to SCCWRP along with the bench sheets, and data quality (i.e., entry error, accuracy, and completeness) was evaluated.

5.5. Data analysis for the baseline intercalibration study

Laboratory performance

A total of 11 laboratories (9 private and 2 public) participated; Lab I could not participate due to microbial contamination causing their culture to crash a few weeks before the start of the ILS. Lab B and M participated in two out of three rounds because the first set of samples was lost during shipping. Lab N reported high mortality in the brood board for one testing round and could only test 3 out of the four samples provided in round 2 (**see Appendix B**). Laboratory performance was considered for individual tests across the three rounds, with an emphasis on laboratory control performance and reference toxicant response.

Like in the historical data, most laboratories met TAC target values for survival and reproduction per surviving female (**Figure 5-5**). Two laboratories did not consistently meet both TAC. Lab B recorded 100% mortality in all their laboratory controls in round 3 whilst > 90% survival was recorded in all the split-samples provided by SCCWRP and tested during that same round. Analysis of the dilution water pointed to errors during the preparation of the dilution water as the calcium to magnesium ratio was inverted (**Appendix B Table B25**). Lab E only met the TAC for reproduction in 50% of their laboratory controls. Lab M and N also had variable neonate production among individual replicates (**Appendix B Table B3**) and met reproduction TAC in ~80% of their laboratory control samples. Lab N indicated that low reproduction was

most likely due to unusually high morality in their test brood board. Approximately half of the laboratories had an average CV for young per female ≤ 0.20 (**Figure 5-6**). Lab A, F and Q reported CVs ≤ 0.15 in at least 10 of their 12 laboratory controls, and Lab G, O and P had CVs ≤ 0.2 in 6 to 9 out of 12 laboratory controls. Conversely, Lab B, E, L, M and N had CVs for reproduction ≥ 0.2 in over 50% of their laboratory controls. Control performance observed in the baseline ILS was consistent with the historical data for most laboratories. Lab A, F and Q continued to produce high quality data, often above established or suggested guidelines. Lab G, O and P showed that they can meet TAC. The CV before reproduction was variable and greater than 0.2, but their overall performance did meet the requirements as described in the EPA manual (2002a). Lab B produced data deemed below expectation based on analysis of both the historical and baseline ILS data. Interestingly, Lab E, M and N exhibited more variable results in the baseline ILS than the patterns observed in their historical data. Laboratory E did not report any unusual circumstances during the ILS, but Lab M had some minor culture issues while Lab N had a major crash of their culture at the beginning of the baseline ILS.

Spiked samples (samples 2B-F and 3) provided further insight into laboratory performance and comparability. These two sample types were used to compare (1) test organism sensitivity to a common reference toxicant (IC25 and IC50, **Appendix B Tables B15 and B16**) and (2) concentration-response patterns and assay precision (PMSD, **Appendix B Table B17**). Lab A, G and Q had PMSD values ≤ 25, with no evidence of bias relative to water type (**Figure 5-7**). Similarly, Lab F, O, and P had PMSD values ≤ 37 in all 6 samples tested regardless of the water type. It should be noted that all these laboratories use different dilution water recipes for their own cultures and test controls **(Appendix A Table A2**). Although this represents a small sample size, the results for these six laboratories are in accordance with the findings of the historical data analysis. Three laboratories had high PMSD above 37 in most samples. Lab N had PMSD > 100 for the SCCWRP-prepared serial dilutions tested in rounds 1 and 2. Laboratories B and M had PMSD values between 44 and 53, and 15 and 64. respectively. These same laboratories had high CVs for reproduction.

To address stakeholders' concerns and evaluate inter-laboratory agreement, statistical analysis of toxicity potency estimates was conducted using the percentile ranking method. The Panel applied this approach for the IC50s and used the interquartile range, the 25th to 75th percentile from the baseline ILS, as lower and upper bounds to characterize data deemed comparable (**Figure 5-8**). Mean IC50 values for six out of 11 laboratories fell within that range, including the high performing laboratories, Laboratories A and Q. There were no obvious differences in the calculated IC50s by water type. For lab F, two out of six IC50s fell within the interquartile range but the laboratory mean was within the 90th percentile upper bound. Laboratories B, E, N and O produced variable (up to 2 order of magnitude difference) IC50s among testing rounds and Lab B and N were deemed not comparable to the other laboratories as the mean IC50s (598 and 757 mg/L NaCl, respectively) fell outside of the 10/90th percentile range. These four laboratories exhibited data characteristics that differ from the level of consistency and comparability characterized in the historical datasets, although Lab B historical IC50s were the least comparable to other IC50s for laboratories using NaCl as their reference toxicant. The percentile ranking approach with the IC50s seems appropriate to identify good, acceptable, and poor laboratory performances. However, larger datasets normally distributed may benefit from tighter upper and lower bounds. While the Panel focused on IC50s that tended to be less variable, data quality programs may consider applying the percentile ranking approach to assess comparability of IC25s or LC50s.

Another metric associated with potency endpoints to consider is the ratio of estimated IC50 to estimated IC25. From a toxicology perspective, this ratio reflects the slope of the concentration-response curve, and it should be consistent within a laboratory, as well as between laboratories for the same toxicant in the same dilution water. Factors that might affect the viability or condition of the test organisms (e.g., a disease challenge, an additional toxicant, changes in water chemistry or test parameters) and otherwise impair their response to a chemical stressor would be expected to increase the degree of response in the most sensitive sublethal indicator, with a concomitant alteration of the concentration-response curve. Conversely, the absence of such a mechanism would be characterized by consistent IC25/IC50 ratios across all labs and waters. This hypothesis was further investigated by comparing IC25/IC50 ratios for laboratories deemed comparable to those of laboratories that had low comparability. Laboratories considered comparable had an average ratio of 0.73, indicating a steep concentration-response curve characterized by a high degree of consistency. Conversely, laboratories that fell outside of the 25^{th} to 75^{th} percentile range had a lower average ratio (< 0.6), as well as greater variability.

Unspiked samples (Samples 1 and 2A) tested during the baseline ILS showed variability in neonate reproduction that was often consistent with the variability in the laboratory's own control samples (**Figures C2-C5**). Laboratories with poor performance also had documented issues with cultures, test brood boards or other technical issues described previously. Further evaluation of the data was conducted using the control adjusted mean reproduction data, calculated as follows: sample mean/control mean x 100 (**Appendix C, Tables C7 and C8**). Using a conservative estimate, control adjusted values greater than 90% were considered not different from the control. The analysis revealed that 4 out of the 11 laboratories participating in the baseline ILS accounted for more than 75% of the instances where the adjusted mean was less than 90%. The adjusted control means were between 14 and 142% for the two samples. The only laboratory reporting toxicity consistently was Lab N, which did not perform well in this exercise. Lab N had high CV for control reproduction, high PMSD and low IC50 comparability (*See analysis of baseline data above).* Lab L had an average adjusted control mean of 90.3% with a range of responses between 76 and 105%. The results suggest that the differences among water types are a function of laboratory performance and not the test method.

Figure 5-5. Mean number of neonates per female in laboratory controls for the baseline and second split-sample exercises. Data is organized by the type of dilution water the laboratory uses in their controls. Laboratories are ordered high to low mean values within each water type. Closed symbols are for the Baseline and Open are for the second ILS. The larger colored symbols are the mean values for each ILS.

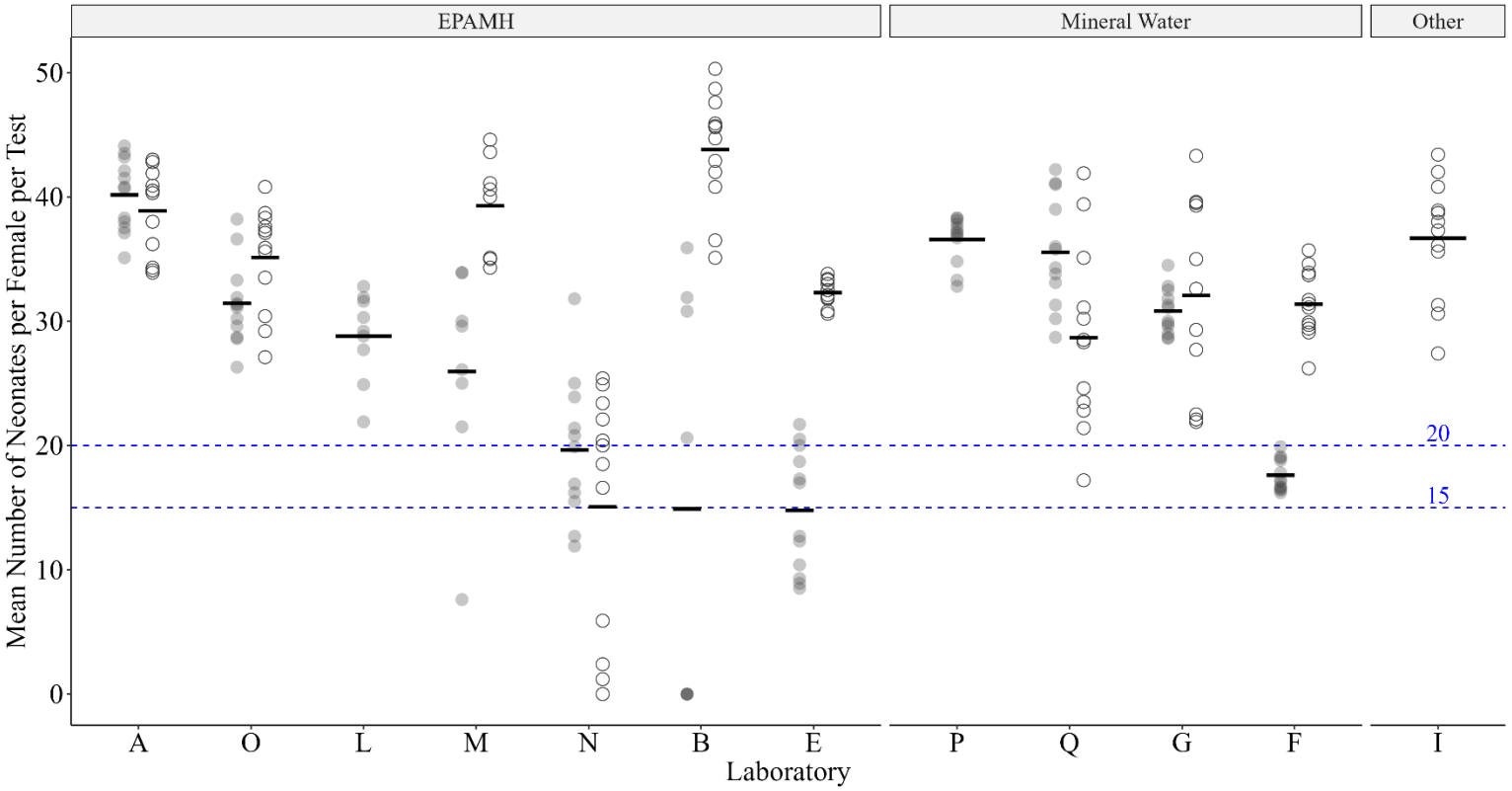


Figure 5-6. Coefficient of variation for the mean number of neonates per female in laboratory controls for the baseline and second split-sample exercises. Data is organized by the type of dilution water the laboratory uses in their controls. Labs are in the same order as mean neonate plot. Closed symbols are for the Baseline and Open are for the second ILS. The larger colored symbols are the mean values for each ILS.

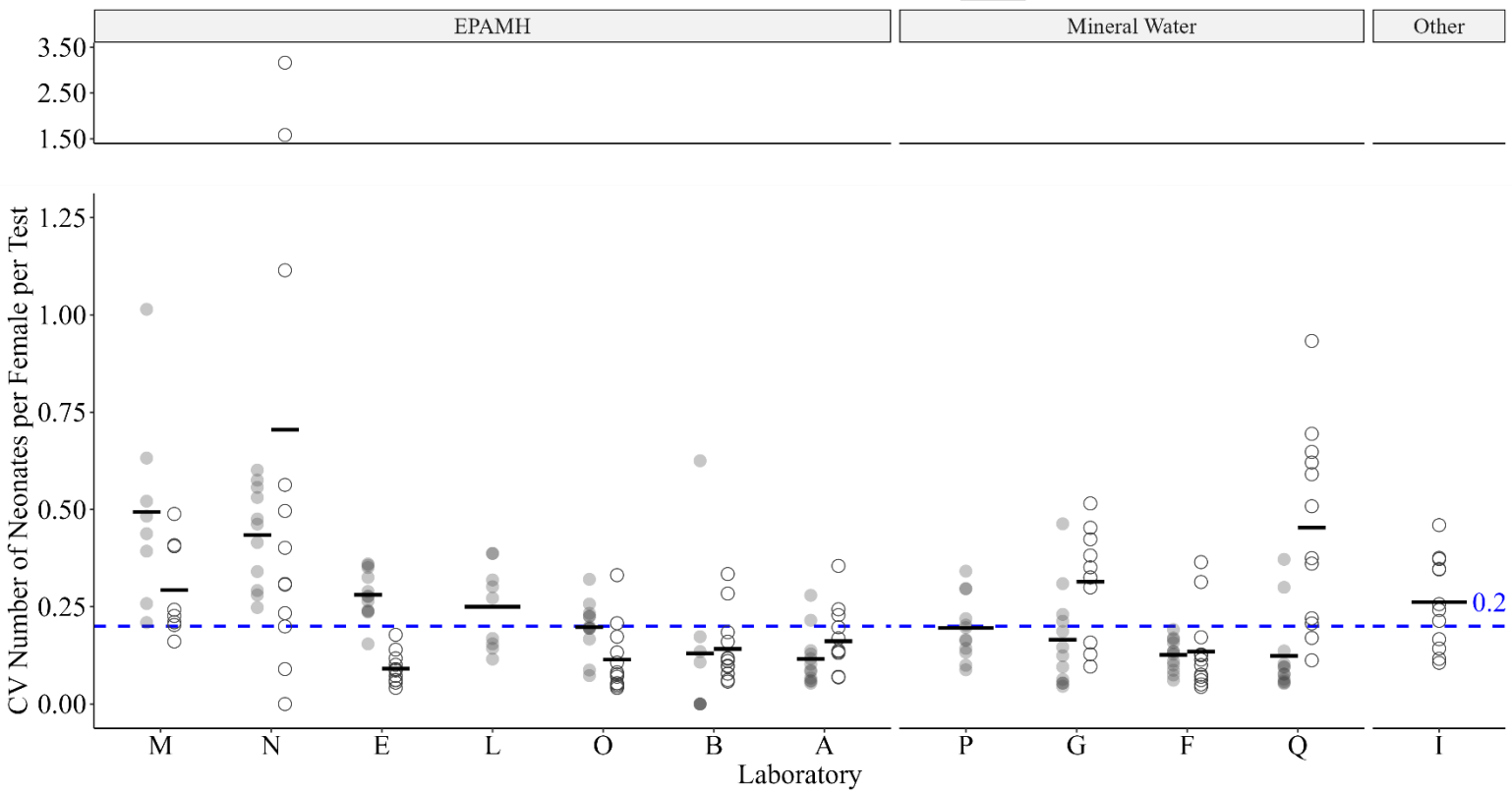


Figure 5-7. Percent minimum significant difference (PMSD) values from both concentration series and both ILS. Reference lines indicate proposed laboratory performance criteria. The dotted lines at 25 and 37 correspond to the 50th and 90th percentile, according to the 2000 EPA study.

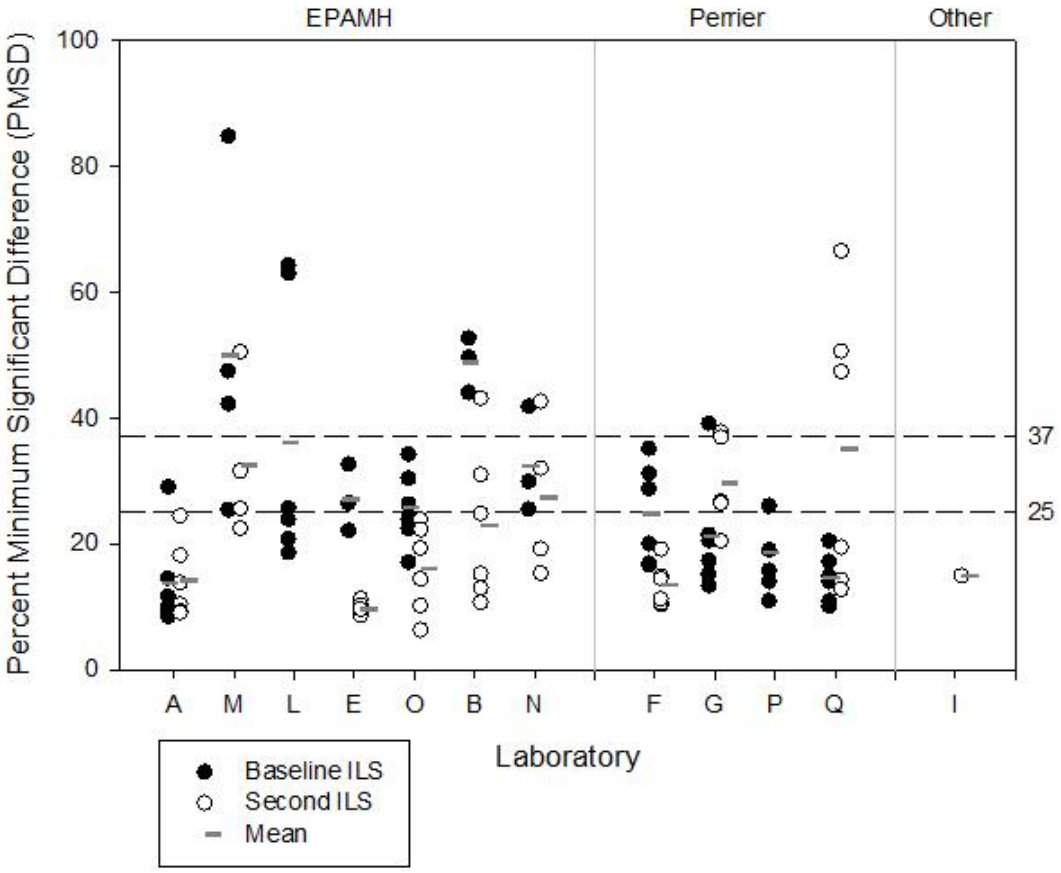
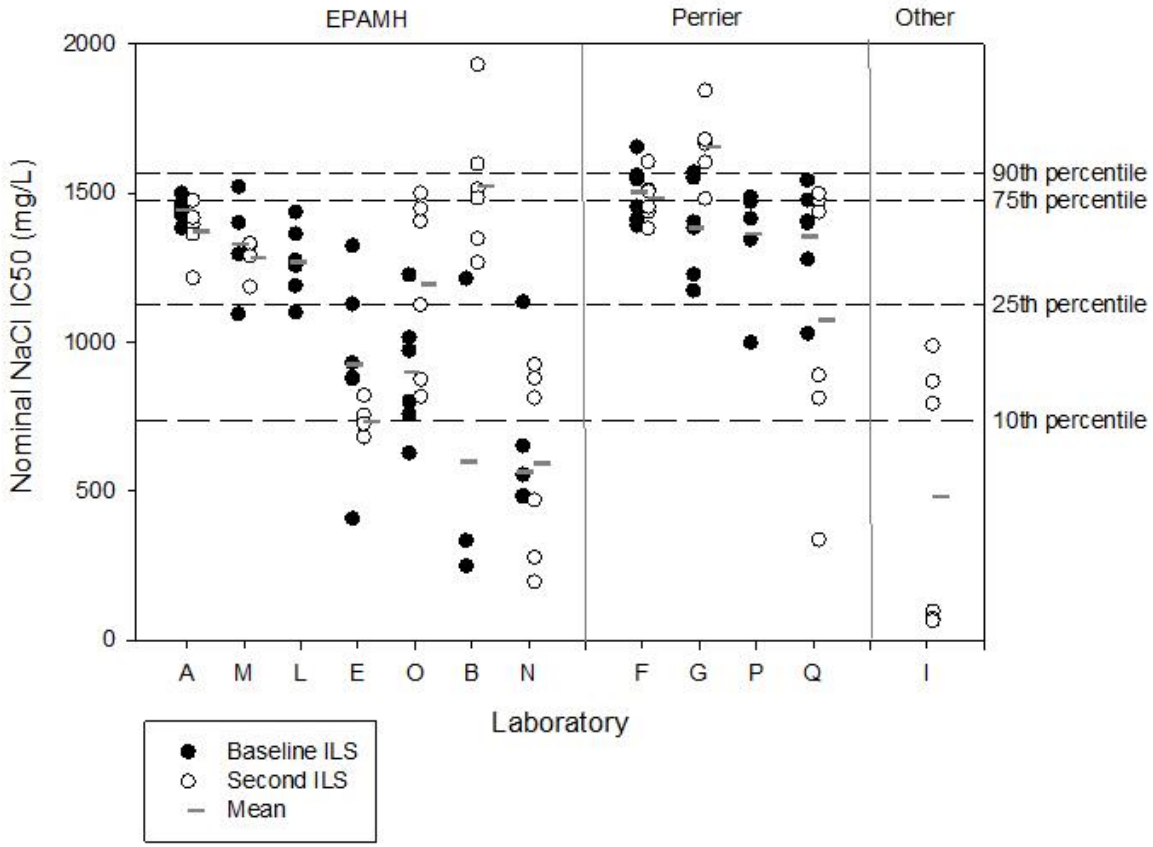


Figure 5-8. Nominal IC50 values from both Sample 2 and Sample 3 concentration series from both ILS. The reference lines are based on percentiles of the entire data set.



Investigating sources of variability in test outcomes

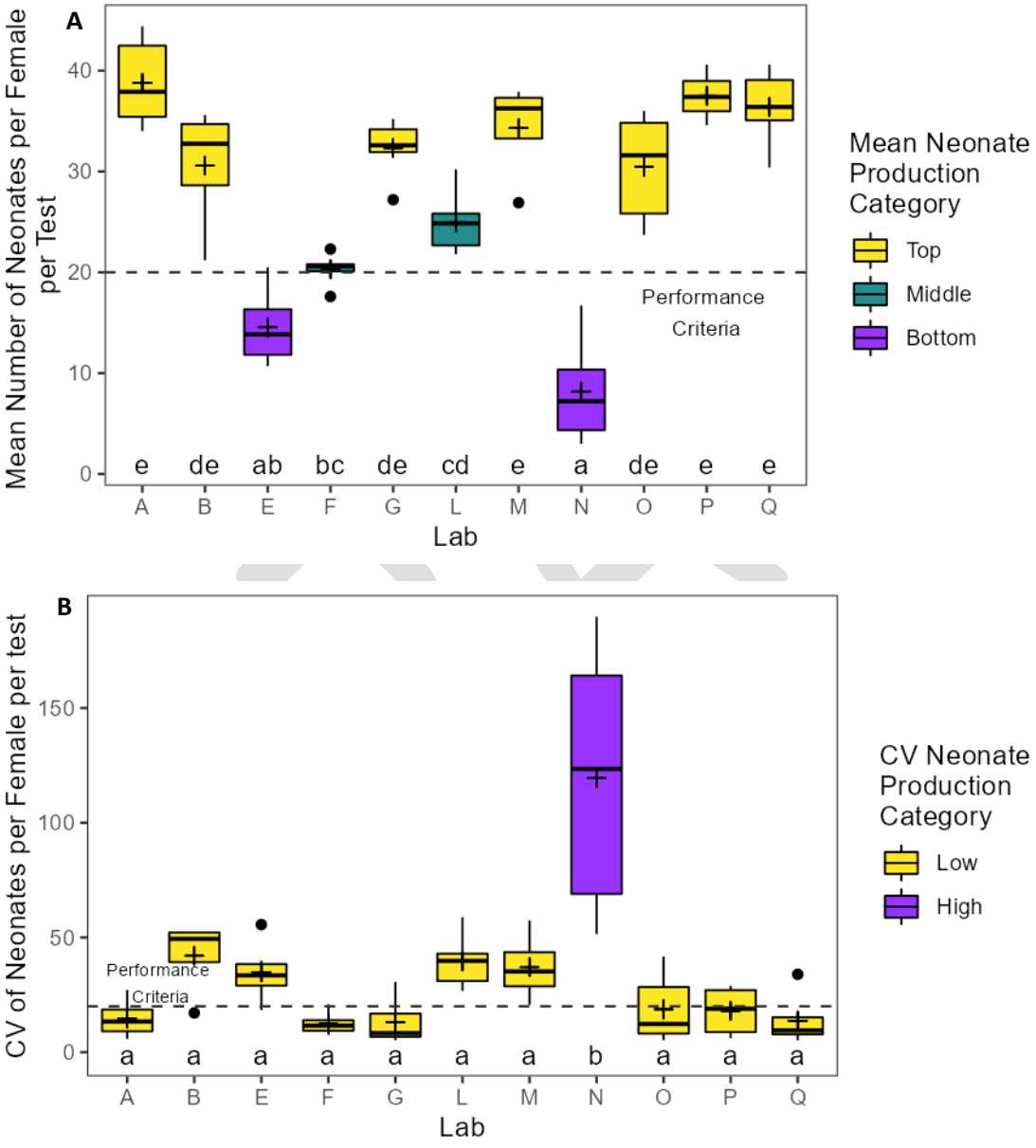
The use of the standardized types of test water helped to reduce the heterogeneity in test results among the laboratories. Patterns in control neonate production of Samples 1 and 2A indicated relatively clear groupings of high performing (i.e., high means and low CVs) and low performing (i.e., low means and high CVs) laboratories (**Figure 5-9**). These groupings were quantified using 2-way ANOVAs using laboratory identity and water type (MHW and DMW) as explanatory variables. Water type was never a significant factor (α=0.05), whereas laboratory identity consistently was a meaningful predictor of both the mean and CV of control neonate production.

Based on the outcomes of the ANOVA analyses, measurements that best captured the techniques and test/husbandry conditions of each laboratory – water composition, test water quality, brood board biological characteristics, brood board water quality, and feeding regimes were investigated as potential influential factors in the differences of neonate production among the laboratories. Variable influence was estimated using random forest regressions, with either mean neonate production or CV of neonate production from Samples 1 and 2A as the response variable. The random forest models indicated that the age of the females in the brood board, brood board water quality (e.g., temperature, pH, conductivity), and test water quality (dissolved oxygen and pH) were important factors in mean neonate production for Samples 1 and 2A (**Appendix B**).

The outputs of the random forest models supported the preliminary conclusions that could be drawn from the lab-to-lab ANOVA analyses, i.e., that practices and conditions inherent to each laboratory were the most likely influences on lab performance. Follow-on linear and logistic regression analyses indicated that the younger the age of the females used to initiate a given test the greater the likelihood of test controls to have higher neonate production with lower variance (**Appendix B, Figures B33-36**). There were also some indications that food quality – specifically the age of the YCT – might be influencing test performance. The linear and logistic models did not indicate that test water quality and brood board water quality were important in explaining test performance.

Taken together, the ANOVA, random forest, and regression model results did not provide a clear-cut indication as to the cause(s) of lab-to-lab variability but helped the Panel and the Stakeholder Committee to refine the laboratory techniques that could be standardized in the second ILS, including age of the females producing the neonates to start the test and feeding regime.

Figure 5-99. Schematic box plots showing differences in mean (A) and CV (B) of control neonate production from Samples 1 and 2A (i.e., MHW and DMW with Perrier®) among the different labs in the baseline intercalibration exercise. The plus symbol indicates the mean value. The letters across the bottom indicate the post-hoc comparisons of each laboratory based upon least-square means Tukey comparisons of the 2-way ANOVA (response=lab\|water type). The colors of the bars indicate groupings based upon visual interpretation and the post-hoc comparisons.



5.6. Data analysis for the second intercalibration study

A total of nine laboratories (8 private and 1 public) participated in the second ILS. Two laboratories, Lab L and P, that participated in the baseline ILS did not participate in this second exercise due to scheduling and staffing issues. One laboratory, Lab N, could only complete two out of three rounds, also due to scheduling and staffing issues. Data were analyzed to evaluate laboratory performance and compare the findings to the baseline ILS and historical data. The goal of this second ILS was to determine whether laboratories producing inconsistent data would show an additional improvement relative to increased control of test procedures. Rather than strict multiple pairwise statistical testing, the Panel evaluated the magnitude of change as a preferable means of determining improvement, if any, between the baseline and second ILS.

All but one laboratory met the TAC for reproduction and recorded ≥ 20 neonates per surviving female. Similarly, all participating laboratories met the TAC for survival, although Lab G often reported less than 100% survival for their controls. Nonetheless, this is a marked improvement from the results of the baseline ILS where three laboratories did not meet one or both TAC (**Figure 5-5**). The calculated CV of young per female also showed a higher frequency of laboratories achieving a CV ≤ 0.2, and Lab B, E and O showed some improvement compared to the baseline ILS (**Figure 5-6**). Lab P exhibited a level of variability that was atypical of their performance in the first ILS and in their historical data. This laboratory did not disclose any issue prior or during testing, or any change in personnel. Consistent with all the data analyzed in this project, Lab A maintained the same reliable level of performance.

As expected, there were no specific standardized techniques that showed a statistically meaningful correlation with the observed improvement. However, implementation of the recommended practices, including consistency in feeding regime through verification of food density in stock bottles and estimation of food in test cup, are known to influence test outcomes, and likely contributed to the overall improvements observed in control performance.

Analysis of the reference toxicant-spiked samples indicated that most laboratories achieved PMSD of ≤ 37, regardless of the dilution water type used (**Figure 5-7**), which provides further evidence that this criterion is not a good indicator of laboratory performance. Half of the laboratories (4 out of 9) had PMSD values ≤ 25 in 5 or 6 out of 6 serial dilutions tested. These four laboratories also met TAC consistently and had average CVs ≤ 0.2. Comparisons of the IC50s showed that data from Labs B, G, I and N fell outside of the 25^{th} to 75^{th} percentile calculated for the second ILS (**Figure 5-8**). Low comparability for Lab G and N was consistent with their poor performance in the laboratory controls, while Lab B did show an improvement relative to the baseline ILS. Lab I had a culture crash during the first ILS and did not participate, then reported 100% mortality in all diluted split-samples provided by SCCWRP in the second ILS. Lab I also reported high mortality in the brood boards used to set the test in the second ILS. It should be noted that this laboratory does not use any of the EPA recipe in the manual for their dilution water.

Analysis of the unspiked samples (Samples 1 and 2A) showed modest improvement in inter-laboratory comparability, although some laboratories remain variable as discussed above. Evaluation of the control adjusted means showed that only two laboratories (B and I) accounted for 75% of the means below 90% **(Appendix C, Tables C7 and C8)**. These laboratories also had documented issues during testing. Similar to the findings from the baseline sample analyses, the variability in response to the unspiked samples is likely due individual laboratory performance.

5.7. Conclusions

Detailed analysis of the historical data and results of the two intercalibration studies showed that several laboratories can perform the *C. dubia* reproduction test and meet test acceptability criteria consistently and exceed expectations for other non-compliance metrics such as CV, PMSD or IC comparability among laboratories. Lab A stood out as one of the most consistent laboratories, and Labs F, O, P and Q were also deemed good performing laboratories based on these metrics of performance. Standardization of laboratory techniques may not improve the performance of these laboratories, but some of the Panel recommendations in the Accreditation and Training categories will improve tracking and documentation of their performance over time. The other laboratories exhibited greater variability as evidenced by the CV for neonate production, potency endpoints and/or PMSD. Interestingly, these laboratories were deemed acceptable based on historical data analysis of laboratory controls, but increased variability was observed during analysis of split-samples. For three of these laboratories, brood board health issues were noted. Two laboratories also reported technical errors impacting dilution water or test maintenance. One laboratory, Lab B, showed clear improvements in performance over the course of this study. Overall, implementation of standardized practices (e.g., detailed brood board health assessment, high quality source water, quantification of food) and performance metrics put forward by the Panel should help improve consistency and ensure that laboratory performance is comparable across California accredited laboratories.

6. Panel's Recommendations to Improve Laboratory Performance and Comparability

The recommended guidance falls into three categories: Best Practices, Accreditation, and Training

6.1. Best Practices

The *C. dubia* EPA promulgated method allows for flexibility in select laboratory techniques to facilitate animal handling and retain necessary flexibility in testing local environmental samples using nationwide guidance. Laboratories can choose the water recipe used to culture the organisms and dilute test samples, food source, age of neonates to start the test (within a 24 h window), etc. This was reflected in the evaluation of the laboratory SOPs and roundtable discussions which revealed that no two California accredited laboratories conduct tests with exactly the same laboratory practices (**Table 5-1** and **Appendix A**). However, there are test parameters that are requirements of the methods which are not flexible, and this study also found that these requirements can be interpreted differently among laboratories. For example, not all California accredited laboratories follow the requirement to end the test after 60% or more females have produced three broods (**Table 6-1**) and some purposely wait for 70% to 80% of the females to produce 3 broods. One laboratory conducts a standard 7-day test regardless of brood status. Such practices could influence toxicity determination when testing environmental samples. To clarify the requirements of the test method and provide additional guidance to improve consistency and comparability, the Panel is providing two sets of guidance. The first set are considered "must do's" and are requirements of the method with the rationale provided in EPA documents. The second set are suggested "should do" recommendations as presented in the EPA manual that may improve laboratory performance.

Below are the "must do's" recommendations.

Recommendation #1: Terminate the test when 60% of surviving females in the controls have had three broods, within a 2-h window (i.e., +/- 1 h of test initiation time).

Termination of the *C. dubia* test when 60% of the females have produced three broods is one of the test acceptability criteria. While the test duration is not specifically prescribed in the method manual, during the Panel's lab visits and data compiled during phone interviews, it was noted that some labs are intentionally delaying test termination later in the day either to allow delayed reproduction in test concentrations to catch up with reproduction in controls, or to accommodate the test breakdown within the lab's schedule (**Table 6-1**). This practice can mask toxicity that is expressed through delayed reproduction. It also produces a source of variability between laboratories, which this study was intended to reduce. For example, analysis of the NaCl-spiked samples tested during the two intercalibration studies shows how test termination trigger can affect the estimated inhibitory concentrations (**Figure 6-1**). As the method states in Section 13.10.9.1, test termination must be completed when 60% of the females or more have produced three broods. Such a decision must be made daily in 24-h increments, within a 2-h window of test initiation time. While the time window is not specifically stated in the method, this is assumed throughout the manual. In multiple instances, "daily" renewal and test "days" are also referred to as 24-h periods. Section 8.5.4 states that the use of samples for static renewal tests be at 24 h, 48 h, and/or 72 h after first use.

Table 6-1.Inventory of laboratory techniques extracted from the laboratory SOPs and phone interviews.

Lab	Water recipe	Feeding method	Test termination; % females having 3 broods	*Test termination window*
A	MHW + vitamins + Se	In test solution	≥60%	*Strict window with single check*
B	Modified MHW	In test solution	≥60%	*Strict window with single check*
C	HW + vitamins + Se	Not provided	≥60%	*No specific window with periodic checks*
D	MHW + Se	Not provided	≥60%	*Strict window with single check*
E	MHW	Direct addition	None	*Test always runs for seven days*
F	80% DIW: 20% Perrier®	Direct addition	≥60%	*No specific window with single check*
G	80% DIW: 20% Perrier®	Direct addition	≥80%	*No specific window with periodic checks*
H	80% DIW: 20% Evian®	Not provided	≥70%	*No specific window with periodic checks*
I	Hoheisel* +vitamins + Se	In test solution	≥60%	*Strict window with single check*
J	Not provided	Not provided	Not provided	Not provided
K	L1650% + vitamins + Se	Not provided	≥60%	*Strict window with periodic checks*
L	MHW + vitamins	Direct addition	≥60%	*Strict window with periodic checks*
M	Modified MHW + vitamins	Direct addition	≥60%	*Strict window with single check*
N	MHW + Se	Direct addition	≥60%	*Strict window with single check*
O	MHW + vitamins + Se	In test solution	≥60%	*No specific window with single check*
P	80% DIW: 20% Perrier®	In test solution	≥60%	*No specific window with periodic checks*
Q	80% DIW: 20% Perrier®	Direct addition	≥60%	*Strict window with single check*

Abbreviations: MHW = EPA moderately hard water; HW = EPA hard water; Se = selenium; DIW = deionized water
*Hoheisel et al. (2011)

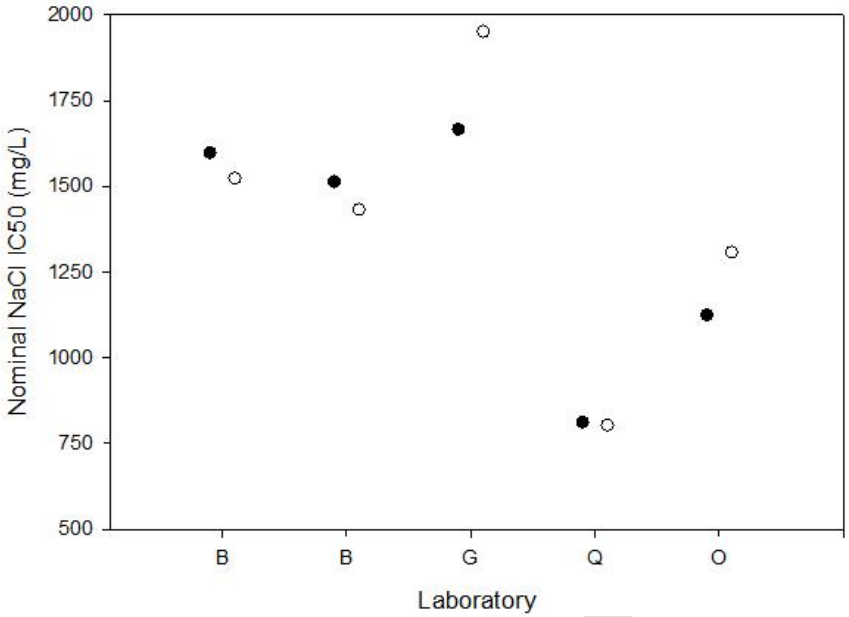


Figure 6-1. Differences in IC50s for tests ended at 60% (●) versus 80% (○) of surviving females having produced three broods. Data was collected during the two intercalibration studies for this project.

Section 13.10.6.1.2 also states that daily routine water chemistry must be measured at the end of each 24-h exposure period. This infers that daily water changes are performed within a reasonably narrow and fixed window each day. It should also be noted that determination of acute endpoints (e.g., lethal concentration at 24h or 48h) requires a precise exposure window (section 13.1.2) which further implies daily checks and test termination at 24-h intervals. Finally, if sufficient toxicity occurs, the 24, 48, and 96 h LC50's could be reported for the test as well.

Recommendation #2: Independently quantify food concentrations in stock bottles and record amounts added to each test container.

The effects of food quality and quantity on *C. dubia* reproduction and toxicity test results are well documented (Cerda and Olive, 1993; Norberg-King and Schimdt, 1993; Jorgenson et al. 2017; Prosser et al. 2018). To support healthy cultures and organisms, the EPA manual provides directives describing the density of YCT and algae to feed daily (see Section 13.10.5). In this study, several practices were uncovered that are not consistent with the manual's directives. Four laboratories indicated not quantifying the density of the stocks used during the test. Interestingly, three of these laboratories (B, E and M) had a significant increase in neonate production in the controls in the second ILS (**Figure 5-5**) which required them to independently quantify the food and estimate the food concentrations in test cups. Other documented practices that are not compatible with the test requirements include filtration of the YCT solution to remove suspended solids. Labs using such technique reported not re-adjusting the density before feeding. To ensure that the density of YCT and algae is within range, laboratories must verify the density of each batch purchased or produced before use and record the volume dispensed to confirm the targeted concentrations in test cups. This is particularly important for laboratories that provide food for the organisms in newly-prepared test solutions instead of adding the food to individual test cups.

Recommendation #3: Use source water produced according to the requirements of the EPA freshwater WET test methods and measure resistance to confirm ongoing water quality.

C. dubia is sensitive to water quality, making it a good indicator species to assess toxicity. To maintain good water quality in the cultures, EPA 2002a provides guidance on the purity of the source water. Section 7.2.2.2 recommends the use of four cartridges to produce deionized source water: (1) ion exchange, (2) ion exchange, (3) carbon, and (4) organic cleanup. This should be followed by a final filtration step. During this study, requests for information on source water treatment systems revealed that laboratories use a variety of water treatment approaches (**Table 6-2**), and few of them are as recommended in the manual. It was noted that most laboratories do not include the filtration step at the end of the treatment train to remove bacteria. Discussions with the laboratories also revealed that most laboratories do not conduct and/or document routine maintenance of their system. Section 5.4.2.1 of the manual requires that laboratories have high quality deionized water providing a resistance of 18 megaohm-cm, also known as Type 1 water (ASTM 1999). However, most of the laboratories visited did not have continuous monitoring of the water resistance to verify that source water consistently meets high standards and did not have a record of the values to refer to. Laboratories must continually or routinely (daily) measure and record the resistance of source water to confirm that it is >18 megaohm-cm and suitable for use in preparing dilution waters.

Recommendation #4: Randomize test cups in the test chamber.

Randomization of test cups on the test boards is a requirement of the test method (section 13.10.2.2) and laboratories must document the set up used for each test. The EPA manual requires the known parentage and provides a procedure for placing test concentration on the board using a stratified random procedure. While most laboratories randomly assign organisms in test cups using blocking by known parentage (section 13.10.2.4), the Science Panel observed that few laboratories randomized the placement of the cups in the exposure chamber. The impact of randomization on test outcome cannot be statistically demonstrated, but randomization is a requirement of the statistical approaches used to analyze test data. This laboratory technique is part of good laboratory practices to avoid any bias due to light, temperature, or other gradients in the exposure chamber. Generating a number randomization templates for the test concentrations can make setting up and transferring the animals easier (cf., EPA 2002a).

Table 6-2. Water treatment methods for laboratories that participated in the study.

Lab	Source Water	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Final Water Type
A	Municipal	Carbon cartridge	Ion exchange	Ion exchange	Filtration	UV disinfection	N/A	1
B	Municipal	Carbon cartridge	Ion exchange	Ion exchange	Other organic clean-up	Other organic clean-up	N/A	1
E	Municipal	Filtration	Ion exchange	Carbon cartridge	N/A	N/A	N/A	1
F	Municipal	Filtration	Carbon cartridge	Reverse osmosis	Ion exchange	N/A	N/A	2
G	Municipal	Filtration	Carbon cartridge	Reverse osmosis	Ion exchange	N/A	N/A	2
I	Municipal	Carbon cartridge	Reverse osmosis	Ion exchange	Filtration	UV disinfection	N/A	1
M	Municipal	Ion exchange	Carbon cartridge	Filtration	UV disinfection	Disinfection: 0.2 µm filter	N/A	1
N	Municipal	Filtration	Carbon cartridge x3	Filtration	Reverse osmosis	Ion exchange	UV disinfection	2+
O	Municipal	Carbon cartridge	Water softener	Filtration	Reverse osmosis	Ion exchange	N/A	1
Q	Arrowhead Distilled Water	N/A	N/A	N/A	N/A	N/A	N/A	N/A

In addition to clarifications of the method requirements, the Panel recommends the following "should do's" best practices. These were developed based on findings of the present study.

Recommendation #5: Conduct a detailed quantitative assessment of brood board health prior to testing.

Documenting brood board health prior to testing is a requirement of the method (13.6.16.11.1). However, the method does not specify the level of detail needed. Initial assessment of the historical data indicated that four out of 17 laboratories consistently documented the health of their brood board in detail, including counts of unhealthy adults and neonates, and the presence of males. Observations recorded by the other laboratories were typically limited to the presence or number of dead organisms. Due to the lack of data collected on brood board health, no meaningful statistical analysis of related historical data could be performed during this study. The Panel noted that three of the five laboratories who did not participate in the intercalibration studies cited issues with their culture prior to testing as a rationale. Therefore, the Panel recommends that laboratories track daily neonate production per female, mortality, number of males, and the size, appearance and movement of adults and neonates n the cultures. Such information provides an ongoing assessment of culture health and serves as an indicator of poor organism quality prior to starting the test. Bench sheets and training materials produced for the second ILS can be used for this purpose (**Appendix D**).

Recommendation #6: Document split brood on bench sheets daily at the time of the observations.

The *C. dubia* test relies on the identification of three separate broods for 60% of the females as an indicator to terminate the test. *C. dubia* has a rapid reproduction rate and can develop and release their brood daily. In some cases, females may release the neonates from a single brood on two consecutive days, which is known as a split brood. The determination of a brood is an important laboratory technique as it directly impacts when the test can be ended and the total number of neonates in three broods. Review of the laboratories' bench sheets revealed that some of them do not add any observations of the females or neonates during daily maintenance. The determination is conducted post hoc by the data analyst, largely based on numbers of neonates. This practice can lead to misidentification of a brood and affect mean neonate production. This was observed firsthand by the Panel during their site visit. In **Figure 6-2**, the top graphic shows a bench sheet of daily neonate counts without any observations, suggesting that each day constitutes a single brood and could underestimate neonate production (see total count for replicate #5 and #6). The bottom graphic shows the same bench sheet with observations of the females' appearance and neonates' size. These observations can be used to identify split broods and better estimate the size of a given brood. Therefore, it is strongly recommended that laboratories make notes of possible split broods on the datasheets at the time of the observations. Bench sheets and training materials produced for the second ILS can be used for this purpose **(Appendix D)**.

		#Neonates/Replicate									
Lab Control	Day	1	2	3	4	5	6	7	8	9	10
	1	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	3	0	0	0
	4	0	0	0	0	2	1	0	1	0	2
	5	4	3	5	0	2	4	3	2	2	2
	6	6	7	6	7	6	7	13	7	12	10
	7	9	10	13	14	14	15	6	10	7	6
	8	7	6	6	5	8	10	3	12	10	12
	Total	19	20	24	21	10	12	22	10	21	14
										Mean	**17.3**

		#Neonates/Replicate									
Lab Control	Day	1	2	3	4	5	6	7	8	9	10
	1	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	3	0	0	0
	4	0	0	0	0	2*	1*	0	1*	0	2*
	5	4	3	5	0	2*	4*	3	2*	2	2*
	6	6	7	6	7	6	7	13	7	12	10
	7	9	10	13	14	14	15	6	10	7	6
	8	7	6	6	5	8	10	3	12	10	12
	Total	19	20	24	21	24	27	22	20	21	20
										Mean	**21.8**

Day 4: Control replicate 6 and 10, females have neonates in pouches.
Day 5: Control replicate 5, 6, 8 and 10, neonates are larger than in the other replicates.

Figure 6-2. Reproduction bench sheet collected from a participating laboratory.

Recommendation #7: Renew test solutions daily within +/- 2 h (i.e., 4-h window) of test initiation time.

As described in Recommendation #1, the *C. dubia* test method is written with the expectation of daily observations and measurements in 24 h increments. Therefore, daily renewal of test solutions should be performed within a consistent window. While this study did not collect data to evaluate the impacts on test outcomes, the practice is consistent with other requirements of the test method, including water chemistry measurements. Further, water chemistries should be measured and recorded on day 0 and on the additional test days as provided in EPA's test methods errata (2016).

Recommendation #8: Upgrade laboratory documentation

Information from historical data collection, laboratory standard operating procedures, and Panel lab visits suggested that routine laboratory operations could be improved with more complete and readily available documentation. This includes documentation of reagent holding times and expiration dates, dilution water preparation and holding times, food preparation and holding times, and randomization templates. Balance logs should be maintained along with calibration records for the balance, and weights that are used to verify the balance is reading correctly. The record book should have a preparation date, the hardness, alkalinity, pH, along with an expiration date. The source of food that is purchased (individual components

or prepared foods) should be recorded in a record book. Sometimes, such information was not properly documented, but more often, information was documented in a format or location that was not readily accessible to laboratory staff that were routinely conducting tests. It is recommended that information about reagent, dilution water, or food holding times be documented directly on the container as well as on data forms that are easily accessible to the laboratory staff. The preparation and use of randomization templates can also make test randomization easier and less error prone.

Recommendation #9: Store reagents to prepare the dilution waters and the reference toxicant appropriately.

The salts used to prepare the EPA synthetic reconstituted water and the reference toxicant NaCl are moisture sensitive. To avoid moisture absorption and changes in the weight of the salts, it is recommended to store them in a desiccator as soon as they arrive in the laboratory. During the laboratory's visits, the Panel noted that these reagents were kept on the bench and not protected from humidity. The recipe for reconstituted water was developed to meet a specific range for water quality parameters, including hardness and alkalinity (section 7.2.3.1 of the EPA manual), and moist salts can impact the ionic composition of the reconstituted water. Similarly, the target ionic balance of the refence toxicant solution can be compromised if the salts used have a higher water content than expected.

6.2. 5.2. Accreditation

California's Environmental Laboratory Accreditation Program (ELAP) has many aspects, but two stand out. The first is Proficiency Testing (PT) where split samples are sent to laboratories. The second is in-person laboratory audits where State staff visit each laboratory and observe lab activities against a checklist of requirements in the promulgated method.

The State of California currently uses the national PT program developed by EPA Discharge Monitoring Report–Quality Assurance (DMR-QA) study program, which consists of laboratories testing a single PT sample annually for the species and method required by the NPDES permit. Yet the procedures for WET are different than for the chemical PT samples. While all PT results are analyzed by the third-party vendor based on one of the toxicity endpoints required by the method: LC50, IC25, NOEC (NOEC-survival, NOEC-growth or NOEC-reproduction), no other PT analyses have multiple reporting values. For the PT program, the unknown sample is purchased from a third-party vendor and tested as a dilution series. Laboratories must ensure that WET test methods/procedures follow instructions from the PT Provider and EPA's promulgated WET test manuals. Laboratories only report one test endpoint (concentration) for each DMR-QA WET test code required. Further, for laboratory performance quality assurance purposes only, the point estimation techniques that produce test endpoints such as IC25 are the preferred statistical approach used for calculating test endpoints for effluent chronic toxicity tests. However, laboratories choose the statistical approach that allows calculation of the test endpoint(s) required by the NPDES permit and the species and method used for routine permit compliance tests. Once the test result is submitted, the third-party vendor is responsible for analyzing and comparing the data to the other laboratories across the country for the same test method. This approach, however, is limited due to the lack of the supporting data generated to report the test result and is a missed opportunity to evaluate other metrics required in EPA method or by

the State, and wide acceptance criteria preventing the detection of important differences in performance across laboratories.

Laboratory audits ensure that test methods are conducted in a consistent manner among laboratories using a defined quality system. For some test procedures, the State may use third party assessors that may not have a strong understanding of the toxicity test methods. For other procedures, the State relies on a single assessor with decades of auditing experience. Evaluation of the accredited laboratory's SOPs during the current study indicated that some deviations from the promulgated method are documented (**Table 6-1**), and site visits of select labs by a subset of the Panel members observed additional inconsistencies with the promulgated method. These findings suggest that audit checklists may not be applied consistently across laboratories and/or that follow-up visits to ensure compliance are not conducted. It should be noted that no Panel member accompanied assessors on an audit, so the inconsistencies are not first-hand observations.

One key concern brought forward by the Stakeholder Advisory Committee is the interlaboratory agreement among ELAP accredited laboratories. Indeed, it is problematic to estimate the level of comparability under the current ELAP program as laboratories can use different dilution waters and reference toxicants in their own QA programs, and can test PT samples from different PT providers. Results of the baseline and second ILS showed that approximately half of the laboratories can produce consistent and comparable results, while the others experienced greater variability than desirable (Figure 6-3).

The following recommendations are largely directed at the State but may also have impact on the laboratories who may need to complete additional testing and reporting, as well as the clients who would potentially absorb some additional costs for the increased accreditation requirements.

***Recommendation #10: Increase number and /or frequency of proficiency testing per year,* following the model used in this study's intercalibration study.**

Currently, laboratories seeking accreditation for the *C. dubia* chronic test are required to participate in one PT study per year. The study consists of testing one sample, with similar test conditions as the NPDES permit, generating one endpoint for each test and then comparing the results to the other laboratories in the nation. Because evaluation of PT results is based on data from the participating laboratories, falling within two standard deviations of the national average allows for more interlaboratory variability than may be acceptable to the State. Moreover, the PT study provides a brief snapshot on laboratory performance but is not sufficient to address intra-laboratory consistency. The intercalibration studies performed in the current project showed that replicate samples provide crucial information on the comparability and repeatability of laboratory results. **Figure 6-3** shows that some laboratories can produce consistent and comparable (most values within 1 SD of the grand mean) results over time (e.g., labs A, F, L), while others are highly variable (e.g., labs B, E, I, N). Note that a similar analysis can be performed using interquartiles for acceptance thresholds (see Section 4).

To demonstrate that ELAP-accredited laboratories can produce comparable results, the Panel recommends analyses of additional PT samples for all laboratories seeking State accreditation. This could be achieved by testing samples from different PT sample providers and analyzing the data using the approach described in section 4 of this document. Test metrics such as control neonate production (mean and CV), IC25 and IC50, concentration response evaluations, and the calculated PSMD would also be evaluated and

compared among laboratories. An alternative to testing more PT samples each year is for all laboratories seeking ELAP accreditation to participate in a bi-annual intercalibration study. A third option for consideration would be to require all labs to use the same reference toxicant, which are typically tested at monthly intervals by accredited laboratories. Each of these three approaches has advantages and disadvantages which should be further discussed and weighed. While they all require additional efforts and costs for the laboratories, such comparisons will better address stakeholders' concerns. The third approach has unique advantages in that laboratories usually already run monthly reference toxicant tests, the data will be immediately available for the laboratory's own use in terms of QA attainment and will offer similar points of comparison across different laboratories for accreditation and procurement purposes.

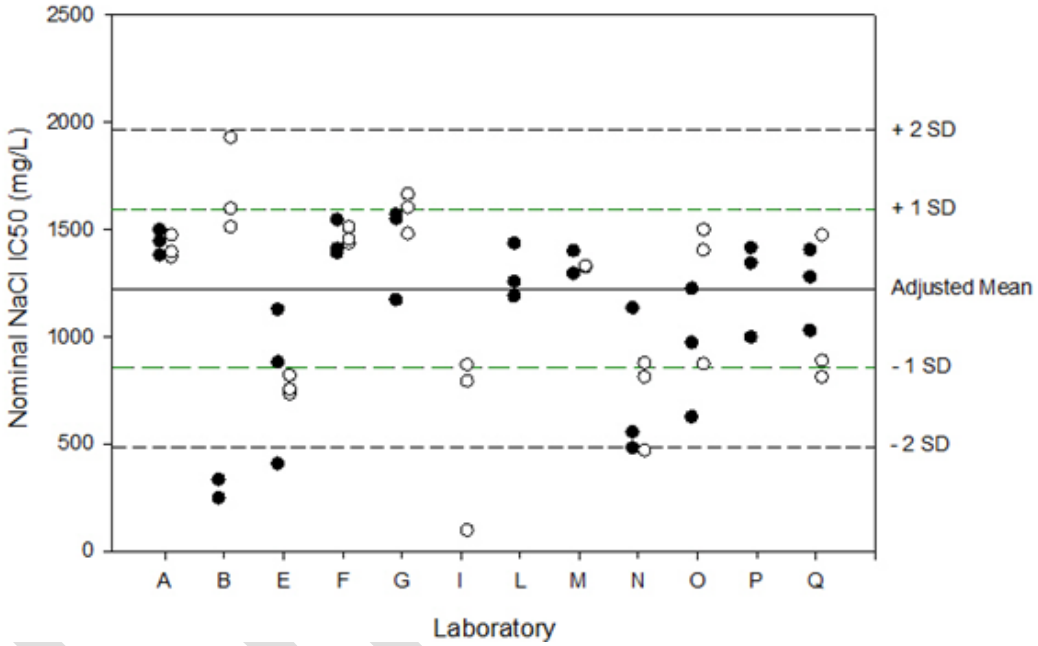


Figure 6-3. Comparisons of IC50s for NaCl-spiked split samples tested during the two intercalibration exercises conducted in Fall 2022 (●) and Spring 2023 (○). All data are plotted (n=59) but only those that met test acceptability criteria (n= 48) were included in the calculations of the mean and standard deviation.

Recommendation #11: Collect and evaluate additional data associated with the PT sample.

Currently, there are three test results evaluated during a PT study: LC50, IC25, NOEC reproduction, NOEC (NOEC-Growth, NOEC-reproduction, NOEC-survival) (**Table 6-3**). These metrics represent only a small fraction of the valuable information that can be collected during PT sample testing. Submission of the data packet with the raw data will allow the evaluation of the mean and CV of neonate production in the controls and in treatment concentrations, the required evaluation of the concentration response, IC25, PMSD and water chemistry. Reviewing these test parameters proved useful during the two intercalibration exercises. The data were used by the Panel to evaluate overall laboratory performance based on their historical data and results of the two intercalibration exercises. Additional information should also be collected to assess replicate level control and treatment responses, brood board health and water quality

(**Table 6-3**). A data framework has already been developed to accommodate these data types through the intercalibrations during the current project and could be valuable in the future. Data submittal templates, data submittal portal, data quality assurance checkers, and pre-programmed data analysis routines have already been designed and utilized and are available to the State. The State will need to host these data submittal requirements and dedicate staff to troubleshoot data submittals and analyze/synthesize the data received.

Table 6-3. C. dubia chronic test endpoints to evaluate as part of proficiency testing.

	Endpoints
Current proficiency testing data	• LC50 • IC25 survival and reproduction • NOEC-survival and NOEC-reproduction
Additional data recommended by the Panel	• Water chemistry daily measurements (initial and final, daily) • Concentration response evaluation • Mean and CV of control neonate production. • IC50 survival and reproduction • PMSD • Ratio IC25/IC50 (reproduction)

Recommendation #12: Optimize laboratory audits to ensure effective and consistent implementation of best practices.

Laboratory audits aim to ensure that accredited laboratories are performing toxicity tests according to the promulgated method requirements and that data quality systems are implemented correctly. Thus, effective audits should be developed to assess critical laboratory techniques (e.g., dilution water preparation, feeding, test termination), and to review data quality procedures in place to maintain laboratory performance goals. This task requires significant time and effort by accreditation officials and is best performed by a trained team of auditors (see Recommendation #13). In the current study, the Panel identified some violations of the test method protocol that can be identified during laboratory audits. To facilitate laboratory audits, the Panel recommends that the State's accreditation checklist be updated to include acceptable practices (e.g., source water treatment, feeding regime, test maintenance) and a list of responses for each type of deviation. There may be other options for optimizing laboratory audits. For example, if there are some issues which are a common thread among laboratories, this could lead to identification of additional "best practices". Another example could be strengthening the review of the corrective actions performed by the laboratories in response to audit findings. Finally, if there are a subset of laboratories which are struggling to maintain their quality standards, these laboratories could be audited more frequently. Formalized approaches will benefit both the State and the laboratories by ensuring that laboratories are equitably and effectively evaluated.

6.3. Training

Although the test method has been in use since 1984, training on the method continues to be needed. EPA has training modules (http://www.epa.gov/water - WET training) and TNI has training on WET as well but additional training may be needed with hands-on type of events. A workshop (EPA Region 9 and California SWRCB) was held just prior to the Expert Panel being convened to provide opportunities for the labs in California to have an open discussion about the *C. dubia* survival and reproduction test method to be able to discuss water conditions, food conditions, ending the test, split broods, data analysis and more. This was held in hopes that labs would be performing the test following the procedures discussed in the workshop. However, as staff changes occur within laboratories, regulators, and regulated entities, continuous access to training is needed to communicate to laboratories, regulators, and the regulated to make appropriate and informed decisions. Roundtable discussions and public meetings with stakeholders highlighted the need to provide training materials for an improved understanding of method requirements and data quality objectives. Training recommendations are directed at the State, the testing laboratories, and at the regulated parties responsible for toxicity testing as a compliance requirement in the following areas:

- Implement training program with defined performance goals for all personnel involved in performing and/or reviewing the *C. dubia* test. This could include how the cultures are managed, the test procedure is performed, collection of effluent samples, reference toxicant testing, data review and analysis, reports that are filed, and PT testing requirements.
- Develop training documentation for standard testing and provide it to the testing laboratories.
- Implement auditors' training program. This type of training would augment auditors' knowledge of the method requirements. The training would aid the development and implementation of training performance goals for laboratory personnel.
- Regulated parties may not have the experience and knowledge to review the data or know how to select laboratories. Training is needed to provide guidance on how to evaluate and review WET test data.

Recommendation #13: Implement auditors' training program.

For effective auditing, State auditors must have a good understanding of the test method including test-specific requirements and general toxicology principles. The State currently has one official auditor responsible for visiting over a dozen labs for accreditation on half a dozen methods. As more auditors become involved, it will be important to ensure that their audits are conducted with similar rigor and attention to detail. A training program would serve to reinforce the key toxicology principles, review and optimize checklists, and develop a forum for ongoing discussion.

Recommendation #14: Implement training program with defined performance goals for all personnel involved in the *C. dubia* test.

As part of the historical data compilation and the two intercalibration studies, the Panel wanted to assess expertise and experience as potential variables leading to intra- and inter-laboratory variability. The assumption was knowledgeable and experienced laboratory staff will produce more consistent and reliable test results. Amongst the California accredited laboratories, anecdotal data compiled during the

laboratory interviews illustrated some labs had long-term staff with decades of experience while others had frequent staff turnover. However, not all labs had a pre-defined training program, few expressed any pre-defined criteria for performance goals, and no labs identified any formal re-testing or continuing education requirements. The greatest challenge, however, was that few of the laboratories documented their training and those that did so, did not conduct or record their training in the same fashion. Ultimately, the Panel could not evaluate the effectiveness of expertise and/or experience for reducing intra- and inter-laboratory variability.

The Panel recommends a training program be created to support the laboratories. The primary objective is to create a training program for each level of activity in the laboratory from bench staff to laboratory managers, and that performance goals are associated with a demonstration of training (and re-training) competence. Moreover, the training should have appropriate documentation which is similar among laboratories so that regulated dischargers can compare training success across potential clients and auditors can utilize the information for troubleshooting accreditation challenges. The training program can be created by the State with assistance of experts in this area who can coordinate the training with a consortium of the testing laboratories, by the regulated dischargers that require well-trained laboratories for compliance assessment, regional trade associations (i.e., Society of Environmental Toxicology and Chemistry), the National Environmental Laboratory Accreditation Program Institute (TNI), or a combination of these groups.

Recommendation #15: Provide guidance to regulated parties to evaluate WET toxicity test data.

The regulated parties on the Stakeholder Committee repeatedly expressed concern about the variability of results within and among California accredited laboratories. However, the Panel noticed that some regulated parties had much more experience and expertise about toxicity testing than others. Some regulated dischargers have their own laboratories and are intimately familiar with the metrics that identify laboratories who produce consistent, high-quality data. In contrast, some regulated parties have only a single compliance officer who is addressing the myriad of facility compliance issues including water, air, and biosolids, amongst others. These compliance officers are typically not aquatic toxicologists and aquatic toxicity is a very small part of their compliance requirements.

The Panel recommends providing guidance to regulated parties on fundamental toxicological concepts, WET testing in specific, and educating compliance assessment staff on the performance metrics that constitute a consistently well-performing laboratory. The goal of this guidance is to empower regulated parties to assess the quality of the laboratories they might wish to retain, and to more fully understand the output of the toxicity tests that determine their compliance. This type of guidance, which can be conducted through a variety of media (written, video, in-person) need not be long or highly detailed, but should be understandable and easily digested by non-toxicologists. Ideally, this guidance should come from a partnership of the regulated dischargers for whom the guidance is directed, but can be done in collaboration with laboratories and/or the State.

7. Limitations

While this study has produced more information on *C. dubia* inter-laboratory variation in more than 15 years, there are still a number of limitations to the conclusions and recommendations provided. These limitations fall into five categories. The first category is the limitation associated with the number of laboratories and the timing of the testing. While the study attempted to generate data for every accredited laboratory in California, 12 out of 17 laboratories participated in the ILS. During this two-year study, at least two laboratories let their accreditation lapse, and at least two laboratories did not provide complete historical datasets. Moreover, for the 12 laboratories who did participate, the magnitude and sources of variability observed during this two-year study may not be similar to variability (or lack of variability) observed prior to or following this study. Implementing the recommendation for increasing the number and/or frequency of proficiency testing samples could help address this limitation in the future.

The second category is the limited capability to quantify the individual variability for each of the nine standardized laboratory best practices. One important finding was the improvement for some laboratories and subsequent reduction in variability from the baseline ILS to the second ILS. The primary difference between the baseline ILS and the second ILS was standardizing the nine laboratory best practices in the recommendations. However, since all nine best practices were changed at the same time, we cannot quantify which best practice provided the most benefit and which provided the least, only the improvement cumulatively across all best practices. To provide this information on the variability associated with each individual best practice, a new study would need to be designed and implemented separating each best practice one at a time.

The third category is the limited capability to quantify variability associated with testing *C. dubia* in dilution water of varying hardness. This study was designed to quantify the variability using the default dilution water in the test method; MHW or DMW. However, some stakeholders wanted to assess the intra- and inter-laboratory variability associated with testing *C. dubia* in HW. The levels of variability may or may not be different in this atypical dilution water hardness. Thus, no recommendation is provided for minimizing variability in the *C. dubia* reproduction test using HW. To address this limitation, a new study would need to be designed and implemented. However, it should be noted the test method does provide guidance for how to control for variability when testing *C. dubia* in HW.

The fourth category is the study's limitation regarding the reference toxicant tested. Quantifying intra- and inter-laboratory variability for the *C. dubia* reproduction test utilized NaCl for concentration-response in both ILS. This concentration-response data was critical in evaluating laboratory performance and consistency. However , the intra- and inter-laboratory variability may differ using other toxicants, especially for those that do not routinely use NaCl. While this element of the study plan was rigorously discussed and evaluated by the Science Panel and the Stakeholders, it was clear insufficient resources and time were available to test additional toxicants.

The last limitation of the study was the timeline. The study's limited timeline and due dates impeded the Science Panel's process. For example, additional time would have provided the opportunity to refine laboratory performance metrics, including developing additional guidance to define consistently

performing laboratories. These activities are encouraged to continue at the conclusion of the study by the State, the regulated dischargers, and the laboratories.

8. References

American Society for Testing and Materials (ASTM). 1999. Standard practice for conducting an interlaboratory study to determine precision of a test method, E691-20. In: Annual Book of ASTM Standards, Vol. 14.05. Philadelphia, PA.

Biau, G. and E. Scornet. 2017. A random forest guided tour. Test 25: 197-227.

Breiman, L. 2001. Random forests. Machine Learning 45: 5-32.

Briden, A.M., S.L. Clark, B.C. Jorgenson, R.S. Ogle, and J. Cotsifas. 2017. *Ceriodaphnia dubia* chronic toxicity test variability: Evaluation of water as a source of test variability. Society of Environmental Toxicology and Chemistry North America 38th Annual Meeting. Minneapolis, MN.

Clark, S.L. and A.M. Briden. 2018. Effect of Ceriodaphnia dubia culture hardness on laboratory control test performance. Society of Environmental Toxicology and Chemistry North America 39th Annual Meeting. Sacramento, CA.

Denton, D. L., Diamond, J., and Zheng, L. 2011. Test of significant toxicity: a statistical application for assessing whether an effluent or site water is truly toxic. Environmental Toxicology and Chemistry 30: 1117-1126.

Denton, D.L., J.F. Fox, J.F., and F.A. Fulk. 2003. Enhancing toxicity test performance by using a statistical criterion. Environmental Toxicology and Chemistry 22: 2323-2328.

Diamond, J., P. Stribling, M. Bowersox, and H. Latimer. 2008. Evaluation of effluent toxicity as an indicator of aquatic life condition in effluent dominated streams: A pilot study. Integrated Environmental Assessment and Management 4:456-470.

Elphick, J.R.F., K.D. Bergh, and H.C. Bailey. 2011. Chronic toxicity of chloride to freshwater species: Effects of harness and implication for water quality guidelines. Environmental Toxicology and Chemistry 30:239-246.

EPA. 2000a. Understanding and accounting for method variability in whole effluent toxicity applications undert the National Pollutant Discharge Eliminatin System Program. Office of Wastewater Management. EPA 833-R-00-003 .

EPA. 2000b. Method Guidance and Recommendations for Whole Effluent Toxicity (WET) Testing (40 CFR Part 136). Office of Water, Washington, DC. EPA 821-B-00-004. July 2000

EPA. 2001a. Final report: Interlaboratory variability study of EPA short-term chronic and acute whole effluent toxicity test methods. Volume 1. Office of Water, Office of Science and Technology. EPA 821-B-01-004.

EPA. 2001b. Final report: Interlaboratory variability study of EPA short-term chronic and acute whole effluent toxicity test methods. Volume 2. Office of Water, Office of Science and Technology. EPA 821-B-01-005.

EPA. 2002a. Short-term methods for estimating the chronic toxicity of effluents and receiving water to freshwater organisms. EPA-821-R02-013. U.S. Environmental Protection Agency. Washington DC.

EPA. 2002b. Guidelines establishing test procedures for the analysis of pollutants: Whole effluent toxicity test methods; Final rule. 67 Fed. Reg. 69952-69972 (November 19, 2002).

EPA. 2002c. Methods for Measuring the Acute Toxicity of Effluents and Receiving Waters to Freshwater and Marine Organisms. 5th edition. Office of Water, Washington, DC. EPA-821-R-02-012.

EPA. 2016. Whole effluent toxicity methods errata sheet. Office of Water. EPA821-R-02-012-ES.

Erickson, R.J., D.R. Mount, T.L. Highland, J.R. Hockett, D.J. Hoff, C.T. Jenson, T.J. Norberg-King, and K.N. Peterson. 2017. The acute toxicity of major ion salts to *Ceriodaphnia dubia*: II. Empirical relationships in binary salt mixtures. Environmental Toxicology and Chemistry 36:1525-1537.

Fox, J.F., D.L. Denton, J. Diamond, and R. Stuber. 2019. Comparison of false-positive rates of 2 hypothesis-test approaches in relation to laboratory toxicity test performance. Environmental Toxicology and Chemistry 38:511-523.

Hoheisel, S.M., R.J. Erickson, T.L. Highland, J.R. Hockett, D.J. Hoff, T.J. Norberg-King, T.W. Valenti, D.R. Mount. 2011. Development and initial evaluation of a reconstituted water formulation that better represents natural waters. Society of Environmental Toxicology and Chemistry North America 32nd Annual Meeting. Boston, MA.

Jorgenson, B.C., S.L. Clark, A.M. Briden, R.S. Ogle, and J. Cotsifas. 2017. *Ceriodaphnia dubia* chronic toxicity variability: Evaluation of food as a source of test variability. Society of Environmental Toxicology and Chemistry North America 38th Annual Meeting. Minneapolis, MN.

Moore, T.F., S.P. Canton, and M. Grimes. 2000. Investigating the incidence of type I errors for chronic whole effluent toxcitiy testing using Ceriodaphnia dubia. Environmental Toxicology and Chemistry 19:118-122.

Mount, D.R., R.J. Erickson, T.L. Highland, J.R. Hockett, D.J. Hoff, C.T. Jenson, T.J. Norberg-King, K.N. Peterson, Z.M. Polaske, and S. Wisniewski. 2016. The acute toxicity of major ion salts to *Ceriodaphnia dubia*: I. Influence of background water chemistry. Environmental Toxicology and Chemistry 35:3039-3057.

Mount, D.R., R.J. Erickson, B.B. Forsman, T.L. Highland, J.R. Hockett, D.J. Hoff, C.T. Jenson, and T.J. Norberg-King. 2019. Chronic Toxicity of Major Ion Salts and Their Mixtures to *Ceriodaphnia dubia.* Environmental Toxicology and Chemistry 38(4): 769–783.

Norberg-King, T.J., and S. Schmidt. 1993. Comparison of effluent toxicity results using *Ceriodaphnia dubia* cultured on several diets. Environmental Toxicology and Chemistry 12:1945-1955.

Prosser, K.N., A.M. Briden, S.L. Clark, and B.C. Jorgenson. 2018. Assessing food preparation and storage as sources of *Ceriodaphnia dubia* culture quality and chronic test variability. Society of Environmental Toxicology and Chemistry North America 39th Annual Meeting. Sacramento, CA.

9. Appendices

9.1. Appendix A – Summary of historical data and laboratory-specific techniques

9.2. Appendix B – Study plan and summary data for the baseline intercalibration study

9.3. Appendix C- Study plan and summary data for the second intercalibration study

9.4. Appendix D – Guidance materials developed during this project to improve documentation of laboratory practices for individual tests.