



UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
WASHINGTON, D.C. 20460

OFFICE OF CHEMICAL SAFETY AND
POLLUTION PREVENTION

MEMORANDUM

DATE: September 27, 2013

SUBJECT: Transmittal of the Meeting Minutes of the FIFRA SAP Meeting Held June 25-28, 2013 on the Scientific Issues Associated with the **“Proposed Endocrine Disruptor Screening Program (EDSP) Tier 2 Ecotoxicity Tests”**

TO: David Dix, Ph.D., Director
Office of Science Coordination and Policy

FROM: Sharlene Matten, Ph.D., Designated Federal Official
FIFRA SAP Staff
Office of Science Coordination and Policy

[Signature] 9/27/13

THRU: Laura Bailey, M.S., Executive Secretary of the FIFRA SAP
Office of Science Coordination and Policy

[Signature] 9/27/13
for

Please find attached the meeting report of the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) Scientific Advisory Panel (SAP) open meeting held in Arlington, Virginia on June 25-28, 2013. This report addresses a set of scientific issues associated with the “Proposed Endocrine Disruptor Screening Program (EDSP) Tier 2 Ecotoxicity Tests”. An electronic copy of this report will be available on the FIFRA SAP website: <http://www.epa.gov/scipoly/sap> and in the public e-docket, <http://www.regulations.gov>, docket: EPA-HQ-OPP-2013-0182.

Attachment

cc:

Jim Jones
Louise Wise
Stephen Bradbury
William Jordan
Margie Fehrenbach
Donald Brady
Susan Lewis
Jack Housenger
Richard Keigwin, Jr.
Robert McNally
Mark Harman
Lois Rossi
Jess Rowland
Oscar Morales
Enesta Jones
Mary Manibusan
Steven Knott
Leslie Touart
Rodney Johnson
Sigmund Degitz
Dale Kemery
Linda Strauss
Molly Hooven
OPP Regulatory Docket

FIFRA SAP Members

Daniel Schlenk, Ph.D. (FIFRA SAP Chair)
Marion Ehrich, Ph.D., DABT, ATS
Stephen Klaine, Ph.D.
Jim McManaman, Ph.D.
Martha Sandy, Ph.D. (FIFRA SAP Session Chair)
Barry Delclos, Ph.D.

FQPA Science Review Board Members

George Bentley, Ph.D.
Vicki Blazer, Ph.D.
Mary Christman, Ph.D.
Robert Denver, Ph.D.
Jennifer Freeman, Ph.D.
Dallas Johnson, Ph.D.
Seth Kullman, Ph.D.
John McCarty, Ph.D.
Edward Perkins, Ph.D.
Catherine Propper, Ph.D.
Colin Scanes, Ph.D., Sc.D.
Kenneth Portier, Ph.D.
Geoffrey Scott, Ph.D.
Shea Tuberty, Ph.D.

SAP Minutes No. 2013-04

**A Set of Scientific Issues Being Considered by the
Environmental Protection Agency Regarding:**

**Proposed Endocrine Disruptor Screening
Program (EDSP) Tier 2 Ecotoxicity Tests**

June 25-28, 2013

FIFRA Scientific Advisory Panel Meeting

Held at

One Potomac Yard

Arlington, Virginia

NOTICE

These meeting minutes have been written as part of the activities of the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA), Scientific Advisory Panel (SAP). The meeting minutes represent the views and recommendations of the FIFRA SAP, not the United States Environmental Protection Agency (Agency). The content of the meeting minutes does not represent information approved by the Agency. The meeting minutes have not been reviewed for approval by the Agency and, hence, the contents of these meeting minutes do not necessarily represent the views and policies of the Agency, nor of other agencies in the Executive Branch of the Federal government, nor does mention of trade names or commercial products constitute a recommendation for use.

The FIFRA SAP is a Federal advisory committee operating in accordance with the Federal Advisory Committee Act and established under the provisions of FIFRA as amended by the Food Quality Protection Act (FQPA) of 1996. The FIFRA SAP provides advice, information, and recommendations to the Agency Administrator on pesticides and pesticide-related issues regarding the impact of regulatory actions on health and the environment. The Panel serves as the primary scientific peer review mechanism of the Environmental Protection Agency (EPA), Office of Pesticide Programs (OPP), and is structured to provide balanced expert assessment of pesticide and pesticide-related matters facing the Agency. FQPA Science Review Board members serve the FIFRA SAP on an *ad hoc* basis to assist in reviews conducted by the FIFRA SAP. Further information about FIFRA SAP reports and activities can be obtained from its website at <http://www.epa.gov/scipoly/sap/> or the OPP Docket at (703) 305-5805. Interested persons are invited to contact Sharlene R. Matten, Ph.D., SAP Designated Federal Official, via e-mail at matten.sharlene@epa.gov.

In preparing these meeting minutes, the Panel carefully considered all information provided and presented by EPA, as well as information presented in public comment. This document addresses the information provided and presented by EPA within the structure of the charge.

TABLE OF CONTENTS

NOTICE	2
INTRODUCTION.....	10
PUBLIC COMMENTERS.....	12
SUMMARY OF PANEL DISCUSSION AND RECOMMENDATIONS	13
OVERALL CONCLUSIONS AND RECOMMENDATIONS	13
GENERAL COMMENTS CONCERNING STATISTICAL ISSUES.....	15
PANEL SUMMARY OF THE JAPANESE QUAIL TWO-GENERATION TEST (JQTT) CHARGE QUESTIONS	23
PANEL SUMMARY OF THE MEDAKA MULTIGENERATION TEST (MMT) AND MEDAKA REPRODUCTION TEST (MRT) CHARGE QUESTIONS.....	30
PANEL SUMMARY OF THE LARVAL AMPHIBIAN GROWTH AND DEVELOPMENT ASSAY (LAGDA) CHARGE QUESTIONS.....	36
PANEL SUMMARY OF THE MYSID TWO-GENERATION TOXICITY TEST (MTTT) AND HARPACTICOID COPEPOD DEVELOPMENT & REPRODUCTION TEST (HCDRT) CHARGE QUESTIONS	41
DETAILED DISCUSSION OF THE CHARGE QUESTIONS.....	49
JAPANESE QUAIL TWO-GENERATION TEST (JQTT) CHARGE QUESTIONS.....	49
<i>Question 1.....</i>	<i>49</i>
<i>Question 2.....</i>	<i>54</i>
<i>Question 3.1.....</i>	<i>55</i>
<i>Question 3.2.....</i>	<i>58</i>
<i>Question 4.....</i>	<i>61</i>
<i>Question 5.....</i>	<i>65</i>
<i>Question 6.....</i>	<i>68</i>
<i>Question 7.....</i>	<i>69</i>
<i>Question 8.....</i>	<i>73</i>
MEDAKA MULTIGENERATION TEST (MMT) AND MEDAKA REPRODUCTION TEST (MRT)	75
<i>Question 1.....</i>	<i>75</i>
<i>Question 2.....</i>	<i>78</i>
<i>Question 3.....</i>	<i>81</i>
<i>Question 4.....</i>	<i>83</i>
<i>Question 5.....</i>	<i>85</i>

<i>Question 6</i>	88
<i>Question 7</i>	91
<i>Question 8</i>	93
<i>Question 9</i>	95
LARVAL AMPHIBIAN GROWTH AND DEVELOPMENT ASSAY (LAGDA)	96
<i>Question 1</i>	96
<i>Question 2</i>	98
<i>Question 3.1</i>	99
<i>Question 3.2</i>	103
<i>Question 3.3</i>	104
<i>Question 4</i>	105
<i>Question 5</i>	109
<i>Question 6</i>	111
<i>Question 7</i>	112
MYSID TWO-GENERATION TOXICITY TEST (MTTT) AND HARPACTICOID COPEPOD DEVELOPMENT & REPRODUCTION TEST (HCDRT) CHARGE QUESTIONS	114
<i>Question 1</i>	114
<i>Question 2</i>	117
<i>Question 3</i>	121
<i>Question 4.1</i>	123
<i>Question 4.2</i>	125
<i>Question 5</i>	127
<i>Question 6.1</i>	128
<i>Question 6.2</i>	130
<i>Question 7</i>	132
<i>Question 8</i>	134
REFERENCES	138
APPENDIX 1: HISTOPATHOLOGY SCORING	147
APPENDIX 2: SUMMARY OF HISTOPATHOLOGY DATA FOR VINCLOZOLIN FROM THREE LABORATORY STUDIES	149

SAP Minutes No. 2013-04

**A Set of Scientific Issues Being Considered by the
Environmental Protection Agency Regarding:**

**Proposed Endocrine Disruptor Screening Program
(EDSP) Tier 2 Ecotoxicity Tests**

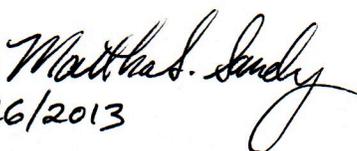
June 25-28, 2013

FIFRA Scientific Advisory Panel Meeting

Held at

**One Potomac Yard
Arlington, Virginia**

**Martha S. Sandy, Ph.D.
FIFRA SAP Session Chair
FIFRA Scientific Advisory Panel**

Signature: 
Date: 9/26/2013

**Sharlene R. Matten, Ph.D.
Designated Federal Official
FIFRA Scientific Advisory Panel
Staff**

Signature: 
Date: 9/26/2013

**Panel Member List for the Meeting of the Federal Insecticide, Fungicide, and Rodenticide Act
Scientific Advisory Panel (FIFRA SAP) to
Consider and Review**

**Proposed Endocrine Disruptor Screening Program (EDSP) Tier 2 Ecotoxicity Tests
June 25-28, 2013**

EPA-HQ-OPP-2013-0812

OPP Docket Tel: 703-305-5805

FIFRA SAP Chair

Daniel Schlenk, Ph.D.
Professor of Aquatic Ecotoxicology & Environmental Toxicology
Department of Environmental Sciences
University of California, Riverside
Riverside, CA

FIFRA SAP Session Chair

Martha S. Sandy, Ph.D., M.P.H.
Supervisory Toxicologist and Chief
Reproductive & Cancer Hazard Assessment Branch
Office of Environmental Health Hazard Assessment
California Environmental Protection Agency
Oakland, CA

Designated Federal Official

Sharlene R. Matten, Ph.D.
U.S. Environmental Protection Agency
Office of Science Coordination & Policy
FIFRA Scientific Advisory Panel
EPA East Building, MC 7201M
1200 Pennsylvania Avenue, NW
Washington, DC 20460
Tel: 202-564-8450, Fax: 202-564-8382
E-mail: matten.sharlene@epa.gov

FIFRA Scientific Advisory Panel Members

K. Barry Delclos, Ph.D.
Research Pharmacologist
Division of Biochemical Toxicology
National Center for Toxicological Research
U.S. Food and Drug Administration
Jefferson, AR

Marion F. Ehrich, Ph.D., D.A.B.T., A.T.S.
Professor of Pharmacology and Toxicology
Department of Biomedical Sciences & Pathobiology
Virginia-Maryland Regional College of Veterinary Medicine
Virginia Polytechnic Institute and State University
Blacksburg, VA

Stephen J. Klaine, Ph.D.
Professor and Director
Clemson University Institute of Environmental Toxicology
Department of Biological Sciences
Pendleton, SC

James L. McManaman, Ph.D.
Professor and Chief
Section of Basic Reproductive Sciences
Departments of Obstetrics and Gynecology & Physiology and Biophysics
University of Colorado-Denver
Aurora, CO

FQPA Science Review Board Members

George Bentley, Ph.D.
Associate Professor
Department of Integrative Biology
University of California - Berkeley
Berkeley, CA

Vicki Blazer, Ph.D.
Research Fish Biologist
Fish Health Branch, Leetown Science Center
U.S. Geological Survey
Kearneysville, WV

Mary C. Christman, Ph.D.
MCC Statistical Consulting, LLC
Gainesville, FL

Robert J. Denver, Ph.D.
Professor
Department of Molecular, Cellular & Developmental Biology
University of Michigan
Ann Arbor, MI

Jennifer Freeman, Ph.D.
Assistant Professor
School of Health Sciences
Purdue University
West Lafayette, IN

Dallas E. Johnson, Ph.D.
Professor Emeritus
Department of Statistics
Kansas State University
Manhattan, KS

Seth Kullman, Ph.D.
Associate Professor
Department of Environmental & Molecular Toxicology
North Carolina State University
Raleigh, NC

John P. McCarty, Ph.D.
Professor of Biology
Director of Environmental Studies
Department of Biology
University of Nebraska at Omaha
Omaha, NE

Reynaldo Patiño, Ph.D.
Unit Leader and Professor
Texas Cooperative Fish & Wildlife Research Unit
U.S. Geological Survey
Texas Tech University
Lubbock, TX

Edward J. Perkins, Ph.D.
Senior Research Scientist
Environmental Networks & Genetic Toxicology
Environmental Laboratory
U.S. Army Engineer Research & Development Center
U.S. Army Corps Engineers
Vicksburg, MS

Catherine R. Propper, Ph.D.
Professor
Department of Biological Sciences
Northern Arizona University
Flagstaff, AZ

Colin G. Scanes, Ph.D., Sc.D.
Department of Animal Science
University of Wisconsin-Milwaukee
Milwaukee, WI

Kenneth M. Portier, Ph.D.
Program Director, Statistics
American Cancer Society
National Home Office
Atlanta, GA

Geoffrey I. Scott, Ph.D.
Director
Center for Coastal Environmental Health
& Biomolecular Research
National Ocean Services
National Centers for Coastal Ocean Science
National Oceanic and Atmospheric Administration
Charleston, SC

Shea Tuberty, Ph.D.
Associate Professor
Department of Invertebrate Physiology
Appalachian State University
Boone, NC

INTRODUCTION

The Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) Scientific Advisory Panel (SAP) has completed its report of the SAP meeting regarding scientific issues associated with the **“Proposed Endocrine Disruptor Screening Program (EDSP) Tier 2 Ecotoxicity Tests.”** Advance notice of the SAP meeting was published in the *Federal Register* (Vol. 78, No. 63) on **April 2, 2013**. The review was conducted in an open Panel meeting on **June 25-28, 2013** at One Potomac Yard, Arlington, Virginia. Materials for this meeting are available in the Office of Pesticide Programs (OPP) public docket or via the public e-docket, **Docket No. EPA-HQ-OPP-2013-0182**, found at the following site: www.regulations.gov). The FIFRA SAP Session Chair, Martha Sandy, Ph.D., chaired the meeting. Sharlene Matten, Ph.D., served as the Designated Federal Official. Steven Knott, Deputy Director, Office of Science Coordination and Policy and Steven Bradbury, Ph.D., Director, Office of Pesticide Programs, provided opening remarks at the meeting. An overview of the EDSP and the proposed EDSP Tier 2 Ecotoxicity Tests was provided by Mary Manibusan, Director, Exposure Assessment Coordination and Policy Division, Office of Science Coordination and Policy. Technical presentations were provided by the following individuals: Leslie Touart, Ph.D., Exposure Assessment Coordination and Policy Division, Office of Science and Coordination Policy; Rodney Johnson, Ph.D., Mid-Continent Ecology Division, National Health and Environmental Effects Research Laboratory, Office of Research and Development and Sigmund Degitz, Ph.D. Mid-Continent Ecology Division, National Health and Environmental Effects Research Laboratory, Office of Research and Development.

Background

Section 408(p) of the Federal Food Drug and Cosmetic Act (FFDCA) requires the EPA to:

Develop a screening program, using appropriate validated test systems and other scientifically relevant information, to determine whether certain substances may have an effect in humans that is similar to an effect produced by a naturally occurring estrogen, or other such endocrine effect as the Administrator may designate [21 U.S.C. 346a(p)].

Subsequent to passage of the Food Quality Protection Act in 1996, which amended the Federal Food, Drug, and Cosmetic Act (FFDCA) and the Federal Fungicide, Insecticide, and Rodenticide Act (FIFRA), and amendments to the Safe Drinking Water Act the same year, the EPA formed the Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC), a federal advisory committee of scientists and stakeholders that was charged by the EPA to provide recommendations on how to implement its Endocrine Disruptor Screening Program (EDSP). The EDSP is described in detail at the following website: <http://www.epa.gov/scipoly/oscpendo/>. Based on the recommendations from the EDSTAC (EDSTAC 1998), the EPA made a number of key decisions using the Administrator’s discretionary authority to include not only the estrogen hormonal pathway, but the androgen and thyroid pathways of the endocrine system in humans as well as in wildlife.

The EDSTAC also recommended the Agency adopt a two-tiered screening and testing program. Tier 1 is an integrated battery of relatively short-term *in vitro* and *in vivo* assays designed to detect the potential of a chemical to interact with the endocrine system, principally the estrogen, androgen, and

thyroid hormonal pathways. Test chemicals determined to have the potential to interact with the endocrine system, based on a weight-of-evidence analysis of the results of Tier 1 screening and inclusive of other scientifically relevant information, would be considered for Tier 2 testing. Tier 2 tests consist of more comprehensive, long-term tests during various life stages and multiple generations enhanced with endocrine-specific endpoints across multiple taxonomic groups, including mammals, birds, fish, amphibians, and invertebrates. The purpose of Tier 2 testing is to identify any potential adverse outcomes and provide quantitative concentration-response information that may be used for risk assessment.

The EDSP is mandated under FFDCA to use “validated” assays to screen and test for endocrine disrupting chemicals. Selected chemicals known to interact with the estrogen, androgen and/or thyroid (EAT) hormonal pathways of the endocrine system were used in the development and validation of each Tier 2 ecotoxicity test according to the general validation principles of the Organization for Economic Co-Operation and Development (OECD) and Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM). The four Tier 2 ecotoxicity tests under consideration are:

- 1) Japanese quail two-generation toxicity test
- 2) Larval Amphibian Growth and Development Assay
- 3) Medaka Multigeneration Test
- 4) Mysid Two-generation Toxicity Test.

The EPA sought the SAP’s advice on factors and results from intra- and inter-laboratory studies that may affect the interpretation of whether or not the four proposed Tier 2 ecotoxicity tests are validated. The Tier 2 tests will be used to provide a more comprehensive assessment of the potential of a test chemical to cause endocrine-mediated adverse effects in the subject taxa. The Integrated Summary Reports (ISRs) and accompanying appendices for each assay provide a detailed account of the historical development and validation of a standardized protocol for each Tier 2 ecotoxicity assay.

PUBLIC COMMENTERS

Oral statements:

- 1) Patricia Bishop, Ph.D., People for the Ethical Treatment of Animals, on behalf of the People for the Ethical Treatment of Animals, the Physicians Committee for Responsible Medicine, and the Humane Society of the United States
- 2) On behalf of the Endocrine Policy Forum
Timothy Fredricks, Ph.D., Monsanto Company;
Katie Coady, Ph.D., The Dow Chemical Company;
Katie Holmes, BASF Corporation;
John M. Brausch, Ph.D, BASF Corporation; and,
Ellen Mihaich, PhD, DABT, Environmental and Regulatory Resources

Written statements:

- 1) Clare Thorpe, Ph.D., CropLife America and Administrator of the Endocrine Policy Forum, on behalf of the Endocrine Policy Forum; and,
- 2) Patricia Bishop, Ph.D., People for the Ethical Treatment of Animals, on behalf of the People for the Ethical Treatment of Animals, the Physicians Committee for Responsible Medicine, and the Humane Society of the United States

SUMMARY OF PANEL DISCUSSION AND RECOMMENDATIONS

Overall Conclusions and Recommendations

The Panel appreciates the hard work EPA has done to implement the 1998 EDSTAC's recommendations on Tier 1 and Tier 2 testing. The focus of this SAP meeting was on the validation status of the proposed EDSP Tier 2 ecotoxicity tests. The Panel supports the scientific rationale and purpose, representative species chosen, biological and toxicological relevance of the major endpoints selected and measured, and the validation process used by EPA for all four proposed Tier 2 ecotoxicity tests. However, the inter-laboratory variability indicates that all four Tier 2 ecotoxicity tests are not yet repeatable and transferable. In particular, there are flaws in the experimental design and statistical analysis, lack of clarity in the protocol and precision in conducting the assays, and a need for additional reference chemicals (generically referred to as positive controls). The Panel provided four general recommendations that will address these deficiencies. The Panel also recommended the development and addition of another avian reproductive test on an altricial passerine bird species, such as the zebra finch, to complement the JQTT. Specific recommendations for each assay are included in the body of the report. The Panel believes that adoption of these recommendations will strengthen the reliability, repeatability and transferability of these assays so they can be used routinely for EDSP chemicals that require Tier 2 ecotoxicity testing.

- 1) Training.** The Panel strongly recommended that laboratory personnel receive specific training for each method to increase proficiency and therefore consistency in the performance of each assay and reduce intra- and inter-laboratory variability. Strict adherence to SOPs, precise measurements of endpoints and interpretation of results are critical. Training areas include: histopathology, analyzing molecular endpoints such as genetic sex for medaka and *Xenopus*, vtg gene expression using qPCR and/or VTG ELISA, and measurement of behavioral endpoints for avian species in particular.
- 2) Statistical methods and experimental design.** The Panel recommended that the experimental design, statistical methods and analysis be discussed in detail in the protocol. A statistical guidance manual for each Tier 2 ecotoxicity test would be very useful. The following aspects should be discussed in detail: randomization, power analysis (sufficient numbers of replicates), and “blinded treatments” including both positive and negative controls to eliminate potential bias.
- 3) Reference chemicals as positive controls.** The Panel strongly recommended that protocols include appropriate positive controls. Positive controls are a fundamental component of good laboratory practices (Fitzpatrick *et al.*, 2008; Kortenkamp *et al.*, 2012; also Munn & Goumenou, 2013; vom Saal *et al.*, 2005; Crofton *et al.*, 2008; Ottinger *et al.* 2008). Use of positive controls would facilitate confidence in the validity of the results by enabling evaluation of assay performance within and among laboratories. Given concerns raised in laboratory performance, experimental design, statistical power, statistical methods and analysis, the Panel indicated that

interpreting negative results from the tests would depend on incorporating appropriate positive controls. Assays conducted without following the Panel's recommendation in this area would be one-tailed tests, where a positive result would provide evidence that a chemical shows endocrine activity, but negative results would not be useful as evidence that a chemical is not an endocrine disruptor. Positive controls will need to be tailored to specific Tier 2 tests because of the different species tested and different endpoints measured in each assay. The choice of positive controls for each Tier 2 test should take into account available information in the scientific literature concerning the signaling pathway(s) that the test can evaluate, e.g., agonist or antagonist activity on estrogen, androgen and thyroid biosynthesis or signal transduction, and the predicted mode of action derived from the Tier 1 assays. Examples include the following:

- estradiol as an estrogen receptor agonist,
- trenbolone as an androgen receptor agonist,
- methimazole as a goitrogenic chemical,
- triiodothyronine (T3) as a thyroid receptor agonist, and
- for invertebrate tests, methoprene, a chemical affecting molting, growth and development in crustaceans,

- 4) **Histopathology studies.** The Panel strongly recommended EPA use a quantitative analysis as a means to enhance analytical power and consistency among laboratories. Where possible, histopathology studies should be conducted using a blind approach coupled with morphometric analysis. A quantitative histopathology analysis would use a systematic series of sections for each tissue conducted using analysis of sections either in a systematic or random basis. In the systematic approach, every 3rd or 5th (or some specific number) section is reviewed. In the random approach, randomization is used to select which sections are reviewed. Image analysis can be readily automated. The Panel indicated that these improvements will reduce bias and facilitate statistical analysis since the data will be continuous and therefore amenable to the use of parametric statistical analysis. This, in turn, will increase consistency in performance and analysis of the histopathology. For the Larval Amphibian Development and Growth Assay (LAGDA), the Panel recommended objective measures such as thyroid area, thyroid follicular cell height, and thyroid follicular cell number using computer-assisted methods.
- 5) **Zebra finch multi-generation avian reproduction test.** The Panel strongly recommended that a multi-generation avian reproduction test be developed (adapt the protocol for JQTT) for zebra finch, an altricial, passerine bird, to supplement the JQTT. An additional test species would be beneficial when bird populations are a specific target of risk assessments or when existing knowledge of environmental exposure indicates that birds are exposed to the chemicals of interest.

General Comments Concerning Statistical Issues

This section summarizes the Panel's general comments concerning statistical issues that apply to charge questions 4, 5, and 6 in all four Tier 2 ecotoxicity assays : (4) *Please comment on the statistical methods chosen for the demonstration and validation of the assays;* (5) *Please comment on the test method robustness and reliability, and the repeatability and reproducibility of the results obtained with the assays and* (6) *Please comment on the transferability across labs and provide any suggestions or recommendations for improvement of the assays.* The Panel's major concern was that the statistical methods and analyses were not detailed enough to thoroughly review them. Additional details should be provided on randomization in all studies, especially for the JQTT and MMT/MRT. Randomization is discussed more appropriately in the mysid multigenerational study and the larval amphibian growth and development study.

The Panel indicated that the overall approach to the statistical analyses fails to provide the information that is desired from a Tier 2 test, namely reproducibility, repeatability, and reliability. Overall, the approaches used in the studies tend to be formulaic and data-driven, are designed to provide conclusions within individual laboratories and fail to provide any approach to comparability across laboratories. The statistical approach used across all four proposed Tier 2 ecotoxicity assays was an analysis of each laboratory's data separately followed by comparison of the test results or coefficients of variation (CV) to determine reproducibility. In this method, a visual comparison of each laboratory's lowest observable effect concentrations (LOECs) or CVs is used subjectively to determine if the laboratories produced consistent results. The Panel recommended that statistical analyses be done independently of the laboratories conducting the assays. The individual laboratory data could be combined into a single statistical model that incorporates laboratory as an effect for each endpoint. This allows for direct testing for inter-laboratory differences and so informs conclusions about reproducibility and robustness.

The Panel identified a second problem with the approach to statistical analysis. Current decisions regarding the statistical method and analysis used are data-driven. As a result, different datasets lead to different analyses. The methods fail to consider expert knowledge and previous research findings that could be used to inform the mode of action (MOA) of the chemical. To understand this problem, the Panel reviewed the statistical methods programmed into StatCHARRMS¹. According to Table 3.3 in the MMT ISR and reproduced below in Table 1, the statistical tests performed by StatCHARRMS depend on whether the dose levels provide normally distributed data with homogeneous (equal) variances. The Panel determined that this assumption is false, i.e., several of the endpoint measures may not be normally distributed and may not have equal variances. This means that an automatic procedure that allows the data to determine the analysis is a concern. Every data set will need to be analyzed using a

¹ StatCHARRMS (Statistical analysis of Chemistry, Histopathology, And Reproduction endpoints including Repeated measures and Multigeneration Studies), a Statistical Analysis System-based program (SAS[®], SAS[®] Institute, Cary, NC) developed for this use by John Green (DuPont Applied Statistics Group, Newark, DE).

different route through the decision tree, which make it difficult to determine what the true Type I error is for this particular method.

Table 1. Decision-tree determining the statistical test used by StatCHARRMS

Table 3-3. Decision-tree determining the statistical test used by StatCHARRMS.

Shapiro-Wilks	Levene's	Statistical test
Pass	Pass	Dunnnett; JT
Fail	Pass	Modified Dunn's; JT
Pass	Fail	Tamhane-Dunnnett; JT
Fail	Fail	JT

Source: Table 3-3, MMT ISR

The result is that each laboratory potentially uses a different statistical method for analyzing the data relying on attributes that are due in part to random variability. Different statistical methods used to analyze the data in each laboratory introduce an additional source of error when comparing data across laboratories and so should be avoided. The Panel stated that the science is what should drive the choice of which statistical model is used rather than an automatic assumption that all data will be normally distributed. Therefore the Panel recommended that scientific expertise in understanding the biology, physiology, biochemistry, ecology, behavior, etc. of the test organisms and in determining the potential toxicological and endocrine-disrupting effects of chemicals in these organisms (and extensions to other organisms) should be used to interpret the distribution of the data and choose the appropriate statistical model to use to analyze the combined data.

The Panel recommended that a complete review of the statistical methods be undertaken. This review should include the data collected in the studies reviewed by the Panel and expert knowledge. The advent of newer statistical methodology and the strong computing capabilities available today allow the best and most appropriate methods be applied to the development of reproducible and robust statistical methods.

The next sections provide a more detailed discussion of the statistical issues and recommendations to improve the statistical analysis.

1) Choice of Probability Distribution

In addition to the problems with the use of a default algorithm to choose the statistical method to be applied, another problem that occasionally occurs is application of inappropriate methods for the data type of the endpoints of interest. The different types of data should be handled distinctly. Continuous variables should be analyzed according to probability distributions that are continuous. The discrete ones should be analyzed according to the probability distribution for discrete random variables take on. And finally, ordinal categorical variables should be handled by categorical techniques, such as the Rao-Scott Cochran-Armitage by Slices (RSCABS) method that has been described by John Green (DuPont), Timothy Springer (Wildlife International), and Amy Saulnier (The Franklin Company) in the

documentation provided (e-docket located at <http://www.regulations.gov>, document identified as EPA-HQ-OPP-2013-0182-0032.pdf). Another alternative is to use models such as cumulative logit models or what are also known as proportional-odds models.

Current statistical methods as instituted in StatCHARRMS appear to treat any quantitative data as continuous and the default is to look for transformations to force data to be normally distributed with homogeneous variance. This causes problems in the review of the outcomes of the Tier 2 testing when determining whether endpoints are appropriate or if the assays have been validated. This introduces further error when comparing results across laboratories because one laboratory may use a very liberal test and another laboratory a very conservative test. As a result, each laboratory could draw a different conclusion. This could explain why, in several of the bioassays, there are differences in the LOECs among laboratories.

The main conclusion from the preceding paragraphs is that decisions as to what statistical models one should use needs to be based on *a priori* knowledge of the types of endpoint measures that are being analyzed. To illustrate, pH is the negative log of the hydrogen ion concentration of a solution and it is typically the measure used to describe the acidity of a solution. It is typically used because pH values tend to be normally distributed; whereas, the reciprocal of the actual hydrogen ion concentration is log-normally distributed. This is a case where a transformation of the data can be performed to create a variable that tends to be normally distributed.

The Panel emphasized that the known science should provide insight into what an appropriate distribution of an endpoint variable is and this is the distribution that should be used when performing statistical tests. This can be accomplished by reviewing the literature for prior research or basing the decision on known likely distributions for a given type of data. For example, count variables are often distributed as a Poisson or Negative Binomial distribution. Whether these distributions fit the data can be checked just as a normal distribution is checked – by comparing the model residuals to the expected quantiles of the proposed distribution (Dunn & Smyth, 1996; Ben & Yohai, 2004). In those cases where the known science has nothing to offer, then one should carefully try to select an appropriate transformation to create a new variable that tends to follow a known probability distribution that one can analyze statistically.

2) Test for Normality

The Panel agreed that among all of the tests for normality, the Shapiro-Wilks test is likely to be the best as it has higher power than the usual alternative tests (Razali & Wah, 2011). However, the Panel suggests that testing for normality may be ill-advised as tests for normality are very much influenced by sample size. With small samples, all normality tests have low power; that is, they will almost always fail to reject normality. In these cases, failure to reject the null hypothesis does not indicate that the null is true. On the other hand, normality tests are overly powerful for large sample sizes, that is, they will almost always reject normality. In fact, distributional tests will likely reject any distribution you choose to test when you have large sample sizes and will fail to reject any distribution that you test when the sample sizes are small. For these reasons, the Panel rejected the assumption that all data, transformed or

not, fit a normal distribution. Instead, the Panel recommended that a visual review of an appropriate graphic, e.g., quantile-quantile plot of the residuals, in addition to a review of the normality test results, should be used to make the determination of whether the data are normally distributed.

The Panel also questioned whether the tests are being applied to the residuals of a fitted model, to the data within each treatment, or to the data combined over all treatments? The Panel recommended that review of normality should be conducted on the model residuals and not the raw data.

3) Homogeneous Versus Heterogeneous Variances

Tests for heterogeneous variances should be performed and Levene's test is an acceptable test that can be used. However, the Panel does not believe that a transformation of the data must be made in order to satisfy a homogeneous variance assumption. Transformations can be used to try to achieve normality, but they need not and should not be used to achieve homogeneous variances as there are statistical packages that can easily analyze data that have unequal variances.

For example, if the data are approximately normal with unequal variances, one can use the MIXED procedure in SAS[®] to analyze the data. Note: SAS[®] commands are written in all CAPITAL LETTERS. The basic commands written for data in which the dose levels have unequal variances and where Y is one of the normally distributed endpoint variables are listed below in Example 1.

Example 1

```
PROC MIXED;  
  CLASS DOSE;  
  MODEL Y=DOSE/DDFM=KR;  
  REPEATED /GROUP=DOSE;  
  LSMEANS DOSE/ADJUST=DUNNETT DIFF=CONTROL('0 dose') CL;  
RUN;
```

The CONTROL option in SAS[®] can be changed to CONTROLU or CONTROLL if one desires one-sided confidence bounds. While the above commands use a REPEATED option they do not require there to be repeated measures being analyzed. If there are repeated measures over time, the SAS[®] commands would be written as shown in Example 2. In this later example, the commands for a repeated measures experiment denote a covariance matrix of the repeated measures to vary by dose level and the experimental units on which the repeated measures occur as ID. Note that TYPE is used to identify the covariance structure of the repeated measures. Possible types of covariance include: CS for compound symmetry, UN for unstructured, and a first order autoregressive (AR1) process with lag 1. Many other types of covariance structures are possible, but the preceding ones are perhaps the most common covariance structures considered.

Example 2

```
PROC MIXED;
  CLASS ID DOSE TIME;
  MODEL Y=DOSE TIME DOSE*TIME / DDFM=KR;
  REPEATED TIME/SUBJECT = ID(DOSE) GROUP=DOSE TYPE=type;
  LSMEANS DOSE/ADJUST=DUNNETT DIFF=CONTROL('0 dose') CL;
  LSMEANS TIME DOSE*TIME/DIFF CL;
RUN;
```

The Panel also suggested the EPA consider the use of the Kenward-Roger (1997) method, DDFM=KR, for adjusting the distributions of the test statistics to account for small sample size. This option should always be used when the model has unequal variances and/or random effects or repeated effects. The Panel added that if there is an interest in combining data across laboratories, for example, when testing whether the same conclusions are reached, similar SAS® commands can be used to address situations where different laboratories have different variances.

The SAS® GLIMMIX procedure can be used to analyze data that do not follow a normal distribution as well as data that follow a normal distribution. Some of the distributions allowed by GLIMMIX are shown in Table 2 below, which is reproduced from the SAS® Help materials. The programming commands are similar to those used in the SAS ® MIXED examples given above. The Panel provided the following references: Milliken & Johnson (2006) for analyses of normally distributed data using the SAS ® MIXED procedure, Stroup (2013) for analyses of non-normal data with the SAS® GLIMMIX procedure and Gbur *et al.* (2012) for detailed descriptions of statistical methods for experiments that yield non-normal data.

Table 2. Keyword values of the DIST=OPTION (Source Table 40.8, SAS ® Online Help & Demonstration Files)

DIST=	Distribution	Default Link Function	Numeric Value
BETA	beta	logit	12
BINARY	binary	logit	4
BINOMIAL BIN B	binomial	logit	3
EXPONENTIAL EXPO	exponential	log	9
GAMMA GAM	gamma	log	5
GAUSSIAN G NORMAL N	normal	identity	1
GEOMETRIC GEOM	geometric	log	8
INVGAUSS IGAUSSIAN IG	inverse Gaussian	inverse squared (power(-2))	6
LOGNORMAL LOGN	lognormal	identity	11
MULTINOMIAL	multinomial	cumulative logit	NA
NEGBINOMIAL NEGBIN NB	negative binomial	log	7
POISSON POI P	Poisson	log	2
TCENTRAL TDIST T	<i>T</i>	identity	10

4) Use of Non-Parametric Tests When Normality Fails

The automatic decision to use a non-parametric alternative to an ANOVA, such as the Kruskal-Wallis Test, when the assumption of normality fails for either the original or transformed data is not necessarily a good one. Nonparametric tests often have even less power to distinguish alternatives than the parametric approaches for the sample sizes used in these assays. The Panel discussed several issues concerning the use of non-parametric tests.

Non-parametric tests have assumptions that must be met just as the parametric tests depend on assumptions. The main distinction is that the non-parametric tests do not specify the shape of the distribution; however, they still rely on equal variances and comparable distribution shapes among levels of the effects. In addition, these tests are quite poor at testing interactions of effects that can be of major importance in the multigenerational assays where it may be important to determine if the F2 generation responds differently than the F1 generation to a treatment.

A rank-based ANOVA, as well as nonparametric analyses, have never been recommended when the underlying assumption of homogeneous variance has been violated, either by itself, or in conjunction with a violation of the assumption of population normality. In general, rank-based and other nonparametric statistics become non-robust with respect to Type I errors for departures from homogeneous variances even more quickly than parametric counterparts that share the same assumption.

The Panel found that the ISRs stated that nonparametric methods can be used when the usual assumption of normality and unequal variances are not satisfied. This was a common misconception repeated throughout the statistical method descriptions in the ISRs. The Panel indicated that there are a few situations, such as one-way treatment designs — either as a completely randomized design or a randomized complete block design, when nonparametric methods are acceptable to use to analyze data that are not normally distributed, but the treatment distributions still need to have approximately equal variances and similar shapes.

The Panel highlighted a problem concerning the use of rank-based nonparametric methods. If an ordinal scale is used to rank the data (e.g., 1, 2, 3, 4, & 5), as done in the histopathology analyses, there will be many observations with the same ranks (a “tie” in other words) and a rank-based nonparametric method will be unable to statistically distinguish them. In this situation, the Panel suggested that it would be better to use categorical methods to analyze such data, e.g., use of RSCABS in histology (see earlier discussion). Another alternative is to use models such as Cumulative Logit Models or what are also known as Proportional-Odds Models.

5) ANOVAs on Rank-Transformed Data

The Panel indicated the validation approaches relied too heavily on nonparametric approaches and made liberal use of performing traditional analyses that assume normality on rank transformed data. There are many cases where these analyses were used in inappropriate situations as discussed below.

Performing ANOVA on ranks means that a standard analysis of variance is calculated on the rank-transformed data. Conducting factorial ANOVA on the ranks of original scores has been suggested by Conover & Iman (1976); Iman (1974) and Iman & Conover (1976). However, Monte Carlo studies performed by Sawilowsky (1985), Sawilowsky *et al.* (1989), Blair *et al.* (1987) and Sawilowsky (1990) and subsequent asymptotic studies by Thompson (1991) and Thompson & Ammann (1989), found that the rank transformation is inappropriate for testing interaction effects in factorial designs. As the number of effects (i.e., main interaction) become non-null, and as the magnitude of the non-null effects increase, there is an increase in Type I error, which will result in a complete failure of the rank transform statistics, with as high as a 100% probability of making a false positive decision. Similarly, Blair & Higgins (1985) found that rank transformation increasingly fails in the two dependent samples layout as the correlation between pretest and post-test scores increase. These findings suggest that applying ANOVA procedures to ranks in experiments involving repeated measures would not be valid.

A variant of rank-transformation is “quantile normalization” in which a further transformation is applied to the ranks such that the resulting values have some defined distribution (often a normal distribution with a specified mean and variance). The quantile normalized data distribution would then be used to compute significance values. The Panel cautioned that two specific types of secondary transformations, the random normal scores and expected normal scores transformation, have been shown to greatly inflate Type I errors and severely reduce statistical power (see Sawilowsky, 1985).

Higgins & Tashtoush (1994) proposed an aligned rank transform test for the problem of testing for interaction in factorial experiments when normal assumptions are violated that addresses the flawed rank transformation procedure discussed earlier in this section. Simulation results have suggested that the test is valid for small to moderate sample sizes when the error distributions are symmetric or moderately skewed. They claim that the procedure has advantages over standard ANOVAs in the presence of outliers or when the error distributions are heavy tailed. They also suggest that the procedure can be extended to repeated measures designs where the repeated measures satisfy compound symmetry.

6) Statistical Power

The Panel stated that determining appropriate sample sizes to achieve an 80% power for detecting a meaningful difference with a Type I error rate of 0.05 is reasonable prior to conducting an experiment and should help ensure that an experiment is worth conducting. However, performing power calculations after an experiment has been completed is not all that useful and is, in fact, an incorrect use of power analyses (Hoenig & Heisey, 2001). The Panel recommended the use of confidence intervals for the mean differences between the control and each of the dose levels since there is an interest in finding the NOECs (No Observable Effect Concentrations) and LOECs of the chemicals being studied. Once an experiment has been conducted, it is much more informative to provide confidence intervals for the differences from the controls at each dose level. The Panel indicated that confidence intervals will provide more information that could be used to characterize the data. Furthermore, if confidence intervals are provided for the true mean differences, then a post-hoc power analysis and coefficients of variation (CV) are not needed.

The Panel indicated that the current designs used in the assays appear to be under-powered for testing the intended hypotheses. The Panel recommended that all test protocols include a section on power analysis where study design issues concerning sample size are discussed in detail prior to the experiment being performed. There are too many toxicology studies that are under-powered for the study objectives. The power analysis would be used to justify the sample size choices. If it is not possible to increase sample size, then alternative approaches should be explored. These range from applying the correct statistical methods to changing the experimental design. Both of these approaches should be based on some analyses of the sources of variability in the data from earlier experiments.

The Panel had some concerns about CVs reported in the protocols. The various protocols had many sources of experimental variation but it was not clear whether these sources of variability were included in the CV calculations. In addition, the CVs were not informative as to whether the protocols were appropriate, powered sufficiently or comparable across laboratories. As a consequence, the Panel recommended clarification of each ISR protocol to discern how a statistically significant effect size is determined.

7) Multiple Comparisons

The description of the test methods for performing the comparison of treatment levels for any of the main effects or their interactions often recommends using Fisher's protected Least Significant Differences (LSD) approach when the overall omnibus F-test rejects the null hypothesis. The Panel indicated that the Fisher's LSD does not, in fact, control the experiment-wise error rate when more than three means are being compared (Einot & Gabriel, 1975); therefore, this approach should be avoided. Alternative methods that control the false positives in multiple testing are preferred and should be based on the desired inferences and particular tests to be performed. This seems to be the case for some of the other tests described in the ISR, but some tests are occasionally recommended under the wrong circumstances. For example, the Jonckheere-Terpstra (J-T) test is recommended if the data display monotonic changes across treatment levels. The Panel stated that the J-T test should not be performed. No test should be done based on the results observed after analyses commence unless the test is designed for sequential review. Otherwise, the false positive rate increases.

Panel Summary of the Japanese Quail Two-Generation Test (JQTT) Charge Questions

Question 1. A rationale for the test method should be available, including a clear statement of scientific basis and the regulatory purpose and need for the test method. The EDSTAC described the Tier 2 tests as having the purpose to “characterize the nature, likelihood, and dose-response relationship of endocrine disruption of estrogen, androgen, and thyroid in humans and wildlife.” Tier 2 tests were designed to be definitive tests that generate sufficient data to characterize the specific hazard of the substance and provide sufficient information on dose-response and adverse effects to permit risk decisions. Please comment on the rationale and purpose of the assay as part of the Tier 2 testing in the EDSP, as described in Sections 2.5 and 2.6 of the JQTT ISR.

Panel Summary

The Panel concluded that the JQTT is scientifically justified and clearly has great potential as a standardized multi-generation reproduction assay for use in the assessment of endocrine disruptor effects on birds; however, the test protocol in its current form needs refinement. The JQTT has a great deal of merit in terms of measurement of the effects of endocrine disrupting chemicals in Galliform birds (Order Galliformes), but it is not clear how broadly applicable the findings will be to other orders of birds with very different reproductive life-histories. For example, phallus size in many ducks is an important reproductive character that is lacking in Galliformes. Perhaps more importantly, in the highly diverse Passerine birds (order Passeriformes), song control system differentiation and size, song behavior, and effects on female choice and sexual behavior could all be affected by endocrine disrupting chemicals used in an agroecosystem such as the Great Plains populated by passerines, and would not be detected by using Japanese quail in this test.

While the data produced using the JQTT will provide a valuable and necessary contribution to inform risk decisions, the Panel did not believe that the JQTT is the “definitive” test that would “provide sufficient information” to conduct an ecological risk assessment for wild birds. The limitations inherent in relying on Japanese quail described here, plus the lack of information about altricial species and many ecologically important endpoints (e.g., parental care, female mating behavior and immune function) support the need for additional data from other bird species and perhaps other tests. Given this situation, the Panel strongly recommended that a multi-generation avian reproduction test be developed on an altricial, passerine bird, e.g., zebra finch (*Taeniopygia guttata*), to complement the JQTT. The Panel provided additional recommendations concerning immune endpoints.

Question 2. Test methods and their associated endpoint(s) should be scientifically relevant to the biological processes of interest and should be demonstrated to be responsive to the specific type of effect/toxicity of interest. Each species presents unique characteristics from a biological perspective and allows for specialized endpoints to address a specific toxicological mode of action. Please comment on the biological and toxicological relevance of the assay in regards to the stated purpose of characterizing endocrine disruptors, as described in Sections 3 and 4 of the JQTT ISR.

Panel Summary

The Panel concluded that the test methods and their associated endpoints for the JQTT are scientifically relevant to the biological processes of interest. The Panel suggested a number of improvements to the protocol that could be incorporated immediately. In addition to the protocol improvements, the Panel strongly encouraged research in the following areas: 1) zebra finch as an additional representative species to complement the Japanese quail and 2) effects of endocrine disrupting chemicals on immune functioning, and potential epigenetic effects (see the detailed responses to charge questions 1 and 7). A detailed discussion of the endpoints is found in the Panel's response to charge question 3.

Question 3.1. The test protocol should be sufficiently detailed and should include a description of what is measured and how it is measured. The selection of endpoints within the assay should be reflective of the biological processes of interest and the endpoints should be intrinsically relevant and have established sensitivity. The test protocol should demonstrate the ability to measure the endpoints and provide adequate performance criteria for evaluation. **Please comment on the selection, optimization and demonstration of the assay endpoints, as outlined in Sections 4 and 5 of the JQTT ISR.**

Panel Summary

The Panel concluded that all of the selected endpoints are scientifically relevant to the biological processes of interest. The Panel concluded that, in general, the JQTT protocol is sufficiently detailed in describing the endpoints measured and how they were measured with some caveats described in the recommendations regarding the selected endpoints. The Panel made the following recommendations.

- 1) **Growth, development, and reproduction.** The Panel recommended that the JQTT continue to include the following endpoints related to growth, development and reproduction: egg production, growth, embryonic development, sex reversal and mortalities.
- 2) **Sexual maturation as an endpoint of endocrine disruption.** The Panel concluded that this endpoint did not add much value and could be dropped from the protocol.
- 3) **Egg shell thickness.** The Panel recommended that egg shell thickness for F0, F1 and F2 continue to be endpoints for the JQTT as it meets the criteria of reliability and relevance.
- 4) **Behavioral indices.** The Panel recommended that behavioral indices be continued as important endpoints in the JQTT.
- 5) **Spermatozoan endpoints.** The Panel recommended that spermatozoa characteristics be included as an endpoint for the JQTT as a decline in spermatozoa number or motility are very likely to influence avian populations.

- 6) **Cloacal area.** The Panel recommended that measurement of the cloacal area is a very good endpoint for endocrine disrupting chemicals with androgenic or anti-androgenic activities.
- 7) **Hormone assays.** The Panel stated that the endpoints chosen by the EPA are broadly appropriate, useful and meet the criteria of reliability and relevance. The CV between different studies is satisfactory. A number of specific recommendations were proposed.
- 8) **Plasma vitellogenin (VTG).** The Panel indicated that the VTG endpoint is not ready for inclusion in the JQTT due to lack of data in the JQTT validation studies measuring the effects of endocrine disrupting chemicals on plasma concentrations of VTG during the validation studies of the JQTT. Therefore, the Panel recommended further research on the effect of endocrine disrupting chemicals on plasma concentrations of VTG and/or hepatic expression of VTG.
- 9) **Epigenetic endpoints.** The Panel recommended the Agency pursue research on possible epigenetic endpoints. Perhaps, there are epigenetic factors that are predictive of reproductive and/or other endocrine modifications that might be worth pursuing.
- 10) **Additional endpoints measured.** The majority of the Panel recommended that endpoint measurements of GnRH-I and –II together with GnIH in the brain, LH, FSH and progesterone in the blood, luteinizing hormone receptor (LHR), follicle-stimulating hormone receptor (FSHR), StAR protein and 17 β -HSD in the gonads, using qPCR, should be included in the protocol. One Panel member did not support including all of these endpoints. Each one would require a great deal of protocol development, assay validation, etc. Another Panel member considered the inclusion of peptide/protein hormones and neuropeptides at least as representative of the physiological mechanisms underlying reproduction as the gonadal steroids.

*Question 3.2. Pathological evaluation of histological tissue preparations is an established, sensitive and integral endpoint in the assessment of effects in long term in vivo assays. The tissues targeted for histopathological evaluation should be shown to be sensitive to exposure and relevant to a mode of action or pathway determination. Tissue samples from several organs or glands (e.g., kidney, liver, thyroid, gonads) underwent histopathological examination in the JQTT inter-laboratory validation and a discussion of histopathology as an endpoint is provided in Section 4.1.13 of the JQTT ISR. **Please comment on the value of histopathological analyses in the JQTT assay for each of the tissue types examined and what (if any) critical information is gained from their inclusion or would be lost if histology were not examined.***

Panel Summary

The Panel concluded that histopathological analyses for all tissue types was well-justified and should be included in the JQTT, i.e., gonadal, epididymal, liver, kidney, and thyroid histopathology. Abnormal histopathology is evidence for endocrine disruption and/or apical effects. The Panel strongly recommended the use of morphometric analysis as a means to enhance analytical power and consistency among laboratories. All treatments should include a positive and negative control with “blind” treatments. The choice of the chemical to be used for a positive control should be made based on the

effects of the putative endocrine disrupting chemical in the molecular/cellular screens/Tier 1 screening assays. The Panel recommended that at least eight random samples (per treatment) for each organ be subjected to histopathological analysis to provide sufficient statistical power. The Panel encouraged research on the applicability of immunocytochemistry/*in situ* hybridization approaches as additional endpoints.

*Question 4. Demonstration of the test method performance should be based on the testing of reference chemicals representative of the types of substances for which the test method will be used. Test substances should adequately represent an appropriate range of responses and physical/chemical properties for which the test method is proposed to be appropriate. The selection of the most appropriate statistical approaches depends in part on the nature of the data and also on the design of the validation study. Statistical and non-statistical methods used to analyze should be described. **Please comment on the selection of test substances and methods (analytical and statistical where appropriate) chosen for the demonstration and validation of the JQTT assay.***

Panel Summary

The Panel indicated that trenbolone was a very informative choice for a test substance, but raised several concerns about the choice of vinclozolin as a central component of the inter-laboratory tests. In particular, the lack of a strong *a priori* expectation of how vinclozolin should impact birds made it difficult to interpret the negative results obtained from the inter-laboratory study. This again emphasizes the need for appropriate positive controls and a formalized system of reference testing for labs to demonstrate proficiency with the protocol. The Panel emphasized that endocrine disrupting substances which introduce biological effects on the mammalian endocrine system might affect the avian endocrine system differently, or not at all. This might be a reason that vinclozolin produced no treatment effects. Use of a different chemical, perhaps trenbolone, as a positive control, would have helped interpretation of the data from the inter-laboratory test

The Panel recommended that a combination of standard laboratory molecular biology tools, e.g., real-time quantitative PCR and immunohistochemistry techniques should be in combination, these techniques will provide information on the endocrine disrupting chemical effects of any test chemical at the molecular, hormonal and structural levels. This thorough analysis, combined with a behavioral analysis, will increase the robustness of the assessment for any test chemical.

The Panel recommended a detailed discussion of how randomization is used in the experimental design be included in the protocol. A detailed discussion of all statistical issues and recommendations for all four Tier 2 assays is found in the section entitled, “General Comments Concerning Statistical Issues”.

Question 5. Considering the variability inherent in biological and chemical test methods, a test method needs to be repeatable and reproducible. A test is robust and reliable if the results are repeatable and reproducible within a laboratory and between different laboratories, respectively. A test protocol should provide sufficient guidance to ensure proper and consistent performance across labs

and chemicals. Please comment on the test method robustness and reliability and the repeatability and reproducibility of the results obtained with the JQTT assay.

Panel Summary

The Panel concluded that several of the tests have high CVs and low power to detect changes; however, there was not enough detail to distinguish between variability introduced by the test protocol and inherent biological variability. If the variability is a function of laboratories using slightly different procedures, then this can be corrected by being more specific in the protocol technical guidance and strict adherence to it. However, if the variability reported reflects biological variation, then the sample size may need to be increased. The Panel provided specific comments and recommendations to improve the test method robustness and reliability and the repeatability and reproducibility of the results obtained with the JQTT assay.

- 1) **Data interpretation and requirement of positive control.** The Panel strongly recommended the use of a positive control chemical with known endocrine disruptor activity should be included in all experiments wherever possible. This would allow for better analysis of the reproducibility and repeatability of the test.
- 2) **Strain differences.** The Panel indicated that using different strains of Japanese quail in the inter-laboratory vinclozolin study is a potential source of variation in the data collected. The Panel recommended that a standard Japanese quail strain be used by all laboratories. This is vital to the repeatability of the assay. The Panel also commented that genetic drift caused by continued colonization in the laboratory might also be a factor causing variation in the inter-laboratory data.
- 3) **Egg storage and incubation parameters.** The Panel stated that some variation in the inter-laboratory results likely arises from subtle differences in humidity, incubation temperature and egg storage temperature employed between laboratories.
- 4) **Time of sampling.** The Panel recommended that a standard time of day of sampling within and between laboratories (mass, hormone measurements) be established.
- 5) **Light intensity.** Housing light intensity should be stated in the protocol and standardized across and within laboratories. All birds should be exposed to the same light intensity in a given photoperiod.
- 6) **Stress lines within a strain and need to measure circulating corticosterone.** Hypothalamic-pituitary-adrenal responsiveness within a strain was not measured to ensure a normal distribution of stress response. The Panel recommended measurement of the corticosterone response to standardized capture-handling stressor to ensure equivalent response across groups and no bias as result of lines with different stress responsiveness. Corticosterone would be more easily measured in serial blood samples and subsequent radioimmunoassay or ELISA, rather than from fecal samples.

- 7) **Histology: tissue selection.** The Panel stated that clear guidance on the selection of animals, tissue, and sections for histological and subsequent statistical analysis should be developed. At present this is vague. The Panel recommended that histological assays (tissue selection, quantitative method, preferably morphometric analysis) be standardized. Tissue sections should be selected on a randomized basis or preferably, a systematic basis, e.g., every 3rd, 5th and 7th section throughout the tissue.
- 8) **Behavioral assays.** The Panel recommended a more detailed and standardized operating procedure be developed to increase transparency in how the assay should be conducted and technical guidance/training to increase the proficiency in performing the assay. Behavioral assays are highly variable and greater standardization and technical guidance would assist in reducing inter-laboratory variability.
- 9) **Sample size.** The Panel recommended that sample size be increased for measurements of all endpoints, particularly hormone sampling, tissue samples and behavioral assays.

Question 6. The test protocol should be descriptive enough to be fully transferable to a competent laboratory. The protocol should describe the methodology of the assay in a clear and concise manner so that a laboratory could comprehend the objective, conduct the assay, observe and measure prescribed endpoints, compile and prepare data for statistical analyses, and report the results. Section 6 of the JQTT ISR outlines the process of and challenges experienced in the inter-laboratory studies. Please comment on the transferability across labs and provide any suggestions or recommendations for improvement of the JQTT assay.

Panel Summary

The Panel stated that the test protocol is not descriptive enough to be fully transferable across laboratories. Data from the inter-laboratory studies are highly variable across many endpoints. Variability among laboratories may be due to poor conduct of the assays, lack of clear and concise descriptions of assay methodology, lack of standard operating procedures, lack of guidance/training for conducting the assays, and choice of statistical methods/analysis. The Panel recommended that a very clear and detailed set of standard operating procedures including instructions on intra- and inter-laboratory standards be more fully developed and included in a revised protocol.

Question 7. The purpose of the validation process is to determine the readiness of a test for inclusion in a testing program. A component of readiness of a test is the evaluation of the usefulness and limitations of the test, including the classes and types of test substances that can and cannot be tested. Please comment on the strengths and/or limitations of the JQTT assay, as described in Section 7 of the JQTT ISR.

Panel Summary

In general, Japanese quail is a solid pragmatic choice as a surrogate bird. Japanese quail have a short generation time and are readily reared in the laboratory. The species has been domesticated and has a long history in captivity. The practical factors that make Japanese quail an attractive species for this test,

also make it more challenging to extrapolate from Japanese quail to wild bird populations that may be the subject of environmental risk assessments. The Panel stated that there were high CVs in the control birds for a number of important endpoints for controls between and within the different laboratories that will limit the power to detect differences. Lack of clarity in the protocol and lack of laboratory proficiency in conducting the JQTT are key sources of variation. The Panel elaborated on other possible sources of intra- and inter-laboratory variation.

The Panel provided the following recommendations in response to this charge question.

- 1) **Defined population of Japanese quail (single source).** There is a critical need for a defined population of Japanese quail available on a widespread (optimally global) basis to test laboratories. The Panel encouraged the Agency to lead efforts to establish, standardize and oversee such a population.
- 2) **A zebra finch two generation reproduction test.** The Panel expressed concern for extrapolating the results of two generation reproduction test using Japanese quail to assessment of the risk to wild birds. As a result, the Panel recommended use of a two generation study with a passerine species, e.g., zebra finch. See the detailed discussion in response to charge question 1.
- 3) **Detail of protocol.** The protocol should provide sufficient detail to ensure that the experimentation is performed correctly and consistently between laboratories.
- 4) **Egg shell thickness and behavioral endpoints.** The effects of putative endocrine disrupting chemicals on egg shell thickness and behavioral endpoints are unique endpoints that should be maintained as part of the JQTT.
- 5) **Epigenetic endpoints.** The Panel recommended that the Agency pursue research on possible epigenetic endpoints. See response to charge question 3.1.
- 6) **Replication and statistical approaches.** The Panel recommended that replication should be increased. Concerns regarding experimental design and statistical approaches are discussed in response to questions 3 and 4, respectively, and in the overview at the beginning of the report.
- 7) **Training programs.** Training programs are critical to understanding the protocol and performance of the JQTT. Laboratories performing the JQTT should have sufficient training such that the laboratories can demonstrate proficiency in the JQTT. A standardized set of training programs (e.g., webinars, web applications, workshops) should be developed and implemented across all laboratories to increase proficiency in running the tests and increase the intra- and inter-laboratory reliability and reproducibility.
- 8) **Clarify terminology.** The use of F0, F1, and F2 terminology is confusing and needs to be clarified. Future documents should clearly state that the F0 generation is the generation in which treatments begin.

Question 8. There is sufficient evidence to indicate that endocrine disrupting chemicals can disrupt normal development and reproductive success, however the sensitivity of the F2 generation compared to

the P0 or F1 is less clearly defined. In the JQTT, there is an increase in endpoint CVs with each subsequent generation (JQTT ISR Table 6-4), which indicates decreased power of discrimination. Please comment from a scientific and risk assessment perspective on the value added of multiple generations in the JQTT assay.

Panel Summary

The Panel recommended maintaining the F2 generation. Multiple generations in the JQTT added significant value to the test results from both scientific and risk assessment perspectives. The F2 generation, where exposure occurs solely during embryonic development, either provides or is capable of providing critical information as to the endocrine disruptor effects of pesticides and other chemicals employed in the environment. Increasing evidence for cross-generation effects emerging from research labs as well as trenbolone results presented in the JQTT ISR suggest that the current JQTT may undervalue the importance of endpoints measured in later generations. Overall, the Panel expressed concern that the statistical power of the current JQTT may be too low. Small sample sizes make it difficult to detect biologically and ecologically important differences, if they exist. The Panel recommended that the number of replicates should be increased in addition to testing multiple generations.

Panel Summary of the Medaka Multigeneration Test (MMT) and Medaka Reproduction Test (MRT) Charge Questions

Question 1. A rationale for the test method should be available, including a clear statement of scientific basis and the regulatory purpose and need for the test method. The EDSTAC introduced the Tier 2 tests as having the purpose to “characterize the nature, likelihood, and dose-response relationship of endocrine disruption of estrogen, androgen, and thyroid in humans and wildlife.” Tier 2 tests were designed to be definitive tests which generate sufficient data to characterize the specific hazard of the substance and provide sufficient information on dose-response and adverse effects to permit risk decisions. Please comment on the rationale and purpose of the assay as part of the Tier 2 testing in the EDSP, as described in Sections 2.1 and 2.2 of the MMT ISR.

Panel Summary

In review of the MMT/MRT ISR (simply MMT ISR), the Panel concluded that each assay provides a clear statement of rationale and purpose. The description in section 2.1 of the MMT ISR is sufficient to justify the need and purpose of the MMT assay, including the necessity for multigenerational testing, full life cycle assessment and appropriate dosing regimen. Overall, the majority of the Panel agreed that clear and concise statements were given regarding the advantages of the medaka model, although there was a recommendation to create a specific list of advantages of using medaka as was done in the JQTT ISR. The Panel recommended that more attention be given to the fact that there is a well-defined mechanism to determine both phenotypic gender and genotypic gender in this model. The Panel stated that this aspect of the model is an advantage for utilization in endocrine disruptor chemical studies that is not seen in other small aquarium fish models. The Panel suggested establishing ratios based on phenotypic gender and genotypic gender (e.g., phenotypic to genotypic ratio) within both the MMT

and/or the MRT multigenerational assays. As currently written, the MMT/MRT protocols did not take advantage of this aspect of the model. Rather, the sole stated purpose for determining genetic sex is to unequivocally identify gender to establish breeding pairs in the F1 and F2 multigenerational study. The Panel addressed each of the components listed in EDSTAC definition of Tier 2 tests, i.e., the nature, likelihood and dose-response relationship of endocrine disruption of EAT² hormonal pathways.

*Question 2. Test methods and their associated endpoint(s) should be scientifically relevant to the biological processes of interest and should be demonstrated to be responsive to the specific type of effect/toxicity of interest. Each species presents unique characteristics from a biological perspective and allows for specialized endpoints to address a specific toxicological mode of action. **Please comment on the biological and toxicological relevance of the assay in regards to the stated purpose of characterizing endocrine disruptors, as described in Section 3 of the MMT.***

Panel Summary

The Panel concluded that the biological and toxicological relevance of the assay is theoretically adequate. The unique characteristics of the medaka life history (as mentioned in response to charge question 1) provide for specific responses throughout the rapid life cycle to address the EAT AOP. These include plasma vitellogenin (VTG), secondary sexual characteristics (SSC) and gonadal histopathology to be evaluated and linked to apical endpoints. The availability of genetic sex markers also allows the evaluation of husbandry conditions which can be refined to provide better reliability and precision. The Panel recommended that an acclimation period be incorporated into the assay time line to establish baseline data for fecundity and fertility prior to F0 exposures.

An additional advantage is that medaka exhibit phenotypic, sexual dimorphic characteristics relevant to use of this species. The Panel recommended that more attention should be given to the fact that there is a well-defined mechanism to determine both phenotypic gender and genotypic gender in this model. The Panel made several additional comments and recommendations that addressed the biological and toxicological relevance of the MMT/MRT assays in regards to the stated purpose of characterizing endocrine disruptors.

*Question 3. The test protocol should be sufficiently detailed and should include a description of what is measured and how it is measured. The selection of endpoints within the assay should be reflective of the biological processes of interest and the endpoints should be intrinsically relevant and have established sensitivity. The test protocol should demonstrate the ability to measure the endpoints and provide adequate performance criteria for evaluation. **Please comment on the selection, optimization and demonstration of the assay endpoints, as outlined in Section 3 of the MMT ISR.***

² EAT = Estrogen, Androgen, and Thyroid hormones

Panel Summary

The Panel agreed that proposed selection of MMT/MRT endpoints exhibit biological and toxicological relevance with regard to endocrine disrupting chemical exposure and assessment. As demonstrated in section 2.2 of the MMT ISR, small aquarium fish, specifically medaka, present unique biological features that make them highly amenable to *in vivo* endocrine disrupting chemical toxicity testing. Numerous endpoints, e.g., growth, mortality, hatching success, and endpoints with a defined HPG (hypothalamic-pituitary-gonadal) axis are based on established EAT AOPs that provide a foundation for the mode of action of defined endocrine disruptor chemicals. These are described as either “activational” or “organizational” endpoints in section 2.2.2 of the MMT ISR.

However, the MMT ISR and provided protocols lack sufficient methodology to fully reproduce this assay. This is particularly true with regards to methodology and data analysis for each defined endpoint of the assay. For example, the Panel recommended the following aspects of the current protocol should be addressed in more detail:

- 1) Water conditions,
- 2) Reproductive activity of fish mating pairs,
- 3) Standardization of VTG analysis,
- 4) Analysis of anal fin papillae,
- 5) Embryo collection standardization,
- 6) Embryo viability,
- 7) Determination of genotypic sex, and
- 8) Hepatotoxicity assays.

*Question 4. Demonstration of the test method performance should be based on the testing of reference chemicals representative of the types of substances for which the test method will be used. Test substances should adequately represent an appropriate range of responses and physical/chemical properties for which the test method is proposed to be appropriate. The selection of the most appropriate statistical approaches depends in part on the nature of the data and also on the design of the validation study. Statistical and non-statistical methods used to analyze should be described. **Please comment on the selection of test substances and methods (analytical and statistical where appropriate) chosen for the demonstration and validation of the MMT and MRT assays.***

Panel Summary

The Panel agreed with the choice of representative chemicals for different endocrine disruption MOAs. The four ER (estrogen receptor) agonists used were: 17 β - estradiol (listed as strong), *o,p'*-DDT (medium), 4-*t*-octylphenol (weak), and 4-chloro-3-methylphenol (weak). Tamoxifen is an ER antagonist and a weak ER agonist. Trenbolone is a strong AR agonist. The EPA stated that prochloraz is an aromatase inhibitor with androgen receptor antagonist activity. However, there is at least one previous study with amphibians that shows prochloraz can also disrupt the thyroid endocrine axis and delay metamorphosis (Brande-Lavridsen *et al.*, 2010). The Panel recommended the Agency add to the description of prochloraz that it also disrupts thyroid-dependent hormone pathways. Vinclozolin is listed

as an AR antagonist and is used in the ring test. However, a recent study with rats presented evidence suggesting that vinclozolin is also a thyroid-active compound that affects thyroid hormone production and/or metabolism (Schneider *et al.*, 2011). The Panel encouraged EPA to review the data for vinclozolin presented in the MMT ISR in the context of potential effects on both the androgen and thyroid endocrine systems.

The Panel stated that Table 3.1 in MMT ISR does a very good job of describing those conditions that can be controlled by the researcher, but a poor job in describing randomization of the fish to these conditions. The Panel raised several questions concerning the experimental design and randomization. For example, the Panel indicated that many details are missing on how to handle the tanks at each generation. Without these details, anyone who tried reproducing the same conditions would have great difficulty. The Panel recommended that the experimental conditions be described much more explicitly. A detailed discussion of all statistical issues and recommendations for all four Tier 2 assays is found in the section entitled, “General Comments Concerning Statistical Issues”.

*Question 5. Considering the variability inherent in biological and chemical test methods, a test method needs to be repeatable and reproducible. A test is robust and reliable if the results are repeatable and reproducible within a laboratory and between different laboratories, respectively. A test protocol should provide sufficient guidance to ensure proper and consistent performance across labs and chemicals. **Please comment on the test method robustness and reliability and the repeatability and reproducibility of the results obtained with the MTT and MRT assays.***

Panel Summary

In conclusion, while the inter-laboratory study between EPA MED and NIES Japan with 4-*t*-octylphenol showed similar data responses between laboratories, the vinclozolin results were much more inconsistent. However, the ring test with vinclozolin is likely more relevant to laboratories that would be conducting the Tier 2 tests in the future. Therefore, for practical purposes, the Panel believed that reliability and robustness has not been fully demonstrated for the MMT. To improve reliability and robustness of the assay, the Panel recommended that the Agency develop more specific and detailed standard operating procedures and guidelines to conduct the assay (see response to charge question 6). In light of the relatively poor performance of the ring test, the Panel suggested that the EPA consider, at least conceptually, the possibility of alternative AOPs for vinclozolin.

*Question 6. The test protocol should be descriptive enough to be fully transferable to a functional laboratory. The protocol should describe the methodology of the assay in a clear and concise manner so that a laboratory could comprehend the objective, conduct the assay, observe and measure prescribed endpoints, compile and prepare data for statistical analyses, and report the results. Section 4 of the MMT ISR outlines the process of and challenges experienced in the inter-laboratory studies and Section 5 presents the optimization and proposed Medaka Reproduction Test (MRT) protocol. **Please comment on the transferability across labs and provide any suggestions or recommendations for improvement of the MMT and MRT assays.***

Panel Summary

The Panel indicated there was modest agreement across MMT trials for a subset of parameters in the inter-laboratory validation studies conducted by EPA MED and NIES Japan with 4-*t*-octyphenol. The Panel indicated that the MMT protocol used by NIES Japan was executed somewhat differently than by EPA MED suggesting that further refinement and clarification should be included to the MMT to ensure global adherence to the same MMT protocol and QA/QC parameters.

One inter-laboratory ring test was conducted between three contract facilities to assess transferability of the MMT. Results from the inter-laboratory ring test were highly variable and suffered from a lack of consistency in both execution of the MMT protocol and overall assay performance. The Panel also raised significant concern regarding the choice of both AOP and chemical agent tested for the initial ring test because the toxicological mechanism of vinclozolin and its metabolites can be complicated and both biochemical and apical endpoints for AR antagonism in medaka are not well defined. As presented, results from the vinclozolin study clearly indicate that the MMT assay, and likely MRT assay, is not currently transferable to other laboratories.

The Panel indicated that there are noticeable components of the MMT and MRT assays that have no documentation that would significantly aid in reducing data variability and enhance overall concordance and transferability of the assay within inter- and/or intra-laboratory trials. While it is expected that most laboratories will have a working knowledge of medaka biology, the Panel recommended detailed protocols be developed to provide sufficient (detailed) guidance for all measurable endpoints. See also the detailed response to charge question 5.

Question 7. The purpose of the validation process is to determine the readiness of a test for inclusion in a testing program. A component of readiness of a test is the evaluation of the usefulness and limitations of the test, including the classes and types of test substances that can and cannot be tested. Please comment on the strengths and/or limitations of the MMT and MRT assays.

Panel Summary

Overall, the Panel identified numerous strengths of the MMT and MRT assay that related to the advantages of using medaka as a model organism, ability to test multiple endocrine disruption AOPs and assay design to assess endpoints in multiple generations.

Some specific advantages of using medaka as a model organism include:

- 1) Husbandry conditions well established
- 2) Ease of manipulation
- 3) Well sequenced genome for reference
- 4) Rich history with available resources in endocrine toxicity
- 5) Short generation time to decrease assay time
- 6) Routine and external spawning with daily spawn
- 7) External embryonic development
- 8) Transparency of the embryos.

- 9) Genotypic sex determination
- 10) Phenotypic sex determination
- 11) Well established and reliable biomarkers of endocrine disrupting chemical effect/exposure (i.e., VTG and SSC)

The Panel also identified numerous limitations of the MMT and MRT assays that were specifically related to experimental design and clarity of the assay protocol. Some of these limitations include:

- 1) Inter-laboratory reproducibility is yet unproven.
- 2) Statistical power of MMT assay is improved with MRT; however, this is at a cost of reduced assessment in F0 and F2 generations, i.e., loss of essential data for AOP and toxicity inference.
- 3) The full potential of phenotypic sex is not utilized, i.e., ratios of phenotypic sex to genotypic sex can be determined.
- 4) There is a lack of proper training and guidance to laboratories to perform the assay.
- 5) There is no assessment of animal health prior to study.
- 6) There is a lack of standardized protocols within the assay for measurement of individual biochemical and apical endpoints.
- 7) There are a limited number of MMT trials with representative chemicals associated with each endocrine disrupting chemical AOP.
- 8) Additionally, there is a need to include reference chemical toxicant that is not mediated by endocrine disrupting chemical effects, such as a reference hepatotoxicant like α -naphthylisothiocyanate.
- 9) There is less certainty with AOPs that represent anti-estrogens and anti-androgens.
- 10) MMT/MRT can be informative of endocrine disrupting chemical AOP, but may be complicated by compounds with putative multiple responses (e.g., tamoxifen can be an ER agonist/antagonist).
- 11) Data related to the effects of steroidogenic disrupting compounds were not as thoroughly evaluated over multiple generations.

*Question 8. There is sufficient evidence to indicate that endocrine disrupting chemicals can disrupt normal development and reproductive success, however the value added of the F2 generation compared to the P0 or F1 is less clearly defined. The EPA is proposing the Medaka Reproduction Test (MRT) as an EDSP Tier 2 assay. The MRT test terminates after the F2 embryo hatch, and a rationale for the proposed protocol is provided in Section 5 of the MMT ISR. **Please comment on the Agency's rationale that the value added by the F2 generation is not sufficient to warrant its inclusion in the Tier 2 fish test protocol.***

The Panel prefaced their response by stating that the data to support the removal of the F2 were from minimal assay trials that had only one comparison of MMT with MRT using 4-*t*-octylphenol. Given the current data set, the Panel concluded that termination of the medaka assay at F2 hatch appears warranted without significant loss of sensitivity for endocrine endpoints. Animal number and labor costs will also be minimized. However, it is recommended that not all F0 endpoints be discarded. In particular,

endpoints that evaluate organ histopathology or overt toxicity (growth) should not be removed due to the potential confounding of non-endocrine maternal effects. With the continuous exposure design in the MRT assay, the Panel expected that results will be similar in F1 and F2. The Panel suggested the addition of tissue chemical analysis to aid in determining if bioaccumulation of the chemical in the tissue is occurring. This would lead to an increase in the overall number of animals for the study. While compounds tested represent a range of endocrine responses, there was still significant uncertainty with AR antagonism and steroidogenesis AOPs.

Question 9. It is the Agency's opinion that the outcomes of the various MMT trials have provided enough information to recommend a medaka reproduction test (MRT) for use as the fish test in Tier 2 of the EDSP. Two major changes from the MMT are proposed, i.e., an increase in the number of replicates per treatment for evaluating effects on reproduction, and terminating the test after the embryos hatch in F2. Other proposed changes include minimizing the collection of endpoint data from F0, and evaluating pathology in only the F1 adults sampled after the assessment of reproduction.

*Overall, the authors conclude that both the MMT and the MRT are transferable methods and are capable of adequately characterizing potential disruption of the endocrine system by putative endocrine disrupting chemicals. However, the MRT is recommended as the preferred EDSP Tier 2 test method for fish because it is less resource intensive with improved statistical power, appears to be as sensitive, and is better able to ensure consistent findings when performed routinely by testing laboratories. **Please comment on the scientific rationale of the Agency's proposed Tier 2 fish test, the MRT, with respect to statistical power, sensitivity, and consistency in performance across laboratories. In addition, please comment on the adequacy of the MRT to characterize potential endocrine disruption, a requirement of Tier 2 of the EDSP.***

Panel Summary

The Panel concluded that the MRT showed clear promise with regard to statistical power, sensitivity and consistency in performance with the 4-*t*-octylphenol test, and the rationale for modification is adequate. However, given the uncertainties inherent in fish husbandry and endocrine variability with this species, additional confirmation with other chemicals is necessary for validation. The 1998 EDSTAC recommended 50-100 compounds go through Tier 1, but it was unclear how many should go through Tier 2. The Panel recommended that representative compounds (both low and high potency) from the ER, ER-A, AR, AR-A, T, T-A and steroidogenesis inhibitor AOPs, as well as a nonendocrine AOP, be evaluated for MRT assay performance prior to adoption as a Tier 2 assay. Ring testing with the MRT (and perhaps one of each AOP) is also recommended.

Panel Summary of the Larval Amphibian Growth and Development Assay (LAGDA) Charge Questions

Question 1. A rationale for the test method should be available, including a clear statement of scientific basis and the regulatory purpose and need for the test method. The EDSTAC introduced the Tier 2 tests as having the purpose to "characterize the nature, likelihood, and dose-response relationship of endocrine disruption of estrogen, androgen, and thyroid in humans and wildlife." Tier 2

tests were designed to be definitive tests which generate sufficient data to characterize the specific hazard of the substance and provide sufficient information on dose-response and adverse effects to permit risk decisions. **Please comment on the rationale and purpose of the assay as part of the Tier 2 testing in the EDSP, as described in Section 3 of the LAGDA ISR.**

Panel Summary

The Panel agreed that the validation methodology used by the EPA is sound and is based on the ICCVAM and OECD published guidelines. The rationale for the assay was well developed in the LAGDA ISR. It includes the ease of monitoring morphological and growth endpoints in amphibians, the ease of exposure to chemicals, and background knowledge and availability of the sensitive endpoints of gonadal and thyroid histomorphology, plasma VTG concentration and possibly plasma tetraiodothyronine (T4) concentration. Furthermore, the ability to both genetically and phenotypically sex the individual animals is important and increases the utility of the assay. All of these endpoints, although subject to overt toxicity, may also indicate outcomes that involve disruption of estrogen, androgen or thyroid hormone production and/or action.

The Panel recommended that a contingency table of expected outcomes for different modes of action for the LAGDA assay be developed as was provided in the ISR for the medaka assay (see Table 3-4, p. 24, MMT ISR). A description of expected outcomes tied to potential MOAs will increase the utility of the LAGDA.

The Panel discussed the issue of extrapolating results of the LAGDA to other amphibian species. Chemical effects (or lack thereof) on *Xenopus* species may or may not be reflective of effects on North American amphibian species. However, the Panel recognized that culture of North American species has not been consistently developed or applied, and therefore, is currently not practical for use in testing protocols. Problems with understanding negative results will continue for environmental risk assessment given the potential for interspecies variation in sensitivity (see the detailed response to charge question 7 for elaboration).

*Question 2. Test methods and their associated endpoint(s) should be scientifically relevant to the biological processes of interest and should be demonstrated to be responsive to the specific type of effect/toxicity of interest. Each species presents unique characteristics from a biological perspective and allows for specialized endpoints to address a specific toxicological mode of action. **Please comment on the biological and toxicological relevance of the assay in regards to the stated purpose of characterizing endocrine disruptors, as described in Sections 3 and 4 of the LAGDA ISR.***

Panel Summary

Overall, the biological and toxicological relevance of the LAGDA to detect developmental abnormalities caused by chemical exposure is strong. Some endpoints may be good measures of endocrine disruption; e.g., thyroid hypertrophy can support a goitrogenic action, elevated plasma VTG can signal an estrogenic action, failure in oviduct formation can signal an androgenic action, and sex reversal may indicate either steroidal or anti-steroid action. However, many effects may not be directly related to endocrine disruption per se, but instead due to a failure in organogenesis, either due to the

chemical influencing specific cellular pathways, or to toxicity. Therefore, distinguishing a non-endocrine from an endocrine MOA will be difficult, and assessing kidney and liver histopathology will be important to help support a specific (possibly endocrine) MOA vs. toxicity. Concerns were raised by the Panel regarding the small sample sizes and the lack of a power analysis for biologically important differences for endpoints that relate to the specific AOPs, although some Panel members thought that incorporating the individual variation into statistical models, as discussed in the section of the report entitled, “General Statistical Comments”, may be helpful.

*Question 3.1. The test protocol should be sufficiently detailed and should include a description of what is measured and how it is measured. The selection of endpoints within the assay should be reflective of the biological processes of interest and the endpoints should be intrinsically relevant and have established sensitivity. The test protocol should demonstrate the ability to measure the endpoints and provide adequate performance criteria for evaluation. **Please comment on the selection, optimization and demonstration of the assay endpoints, as outlined in Sections 3 and 4 of the LAGDA ISR.***

Panel Summary

The Panel indicated that the selected endpoints are appropriate for evaluating chemical effects on growth and development of *Xenopus* larvae, and on gonad and thyroid gland, and urogenital tract organogenesis that may reflect endocrine disruption. Alterations in growth of larvae combined with clear liver and possibly also kidney histopathology may be scored as toxicity. Effects on growth, particularly at lower concentrations of test chemicals, and possibly without clear liver, or kidney, histopathology could indicate disruption of specific cellular signaling pathways (endocrine or non-endocrine).

*Question 3.2. Three of the four labs in the inter-laboratory validation of the LAGDA were unsuccessful in properly performing the thyroxine (T4) analyses using commercially available Enzyme-linked Immunosorbent assay (ELISA) kits with antibodies specific for human or canine T4). Subsequently, the EPA developed an extraction method (Section 5.3.2) and presents a revised ELISA method for measuring T4 in amphibian samples in Appendix 7 of the LAGDA ISR. **Please comment on the technical feasibility, reproducibility, and accuracy of the revised ELISA T4 measurement method. Provide any recommendations regarding additional guidance to ensure the reproducibility of T4 measurements.***

Panel Summary

Given the technical challenges in measuring plasma T4 and the limited interpretative power of measurements at the single stage of metamorphosis, NF stage 62, the Panel recommended that the EPA consider eliminating plasma T4 measures from the LAGDA.

*Question 3.3. Vitellogenin (VTG) is an established biomarker of estrogenic exposure and is used as a key endpoint in endocrine disruptor testing. There is currently not a standardized commercial source for *Xenopus laevis* VTG antibodies for an ELISA. **Please comment on the protocol for measuring and reporting VTG levels. Provide any recommendations regarding additional guidance to ensure the consistency, repeatability and reproducibility of VTG measurements.***

Panel Summary

Plasma VTG or liver *vtg* mRNA measurements are useful diagnostic tools for disruption of sex steroid signaling, particularly estrogenic signaling. There was large variation in the control plasma VTG concentration measurements among the contract laboratories that conducted the assays. The reasons for the variation are unknown. Most were unable to detect VTG and this is likely due to the immaturity of the frogs at 10 weeks post-metamorphosis. There are two issues to be addressed: 1) the performance of the VTG ELISA and 2) the inability to detect plasma VTG in immature frogs.

*Question 4. Demonstration of the test method performance should be based on the testing of reference chemicals representative of the types of substances for which the test method will be used. Test substances should adequately represent an appropriate range of responses and physical/chemical properties for which the test method is proposed to be appropriate. The selection of the most appropriate statistical approaches depends in part on the nature of the data and also on the design of the validation study. Statistical and non-statistical methods used to analyze should be described. **Please comment on the selection of test substances and methods (analytical and statistical where appropriate) chosen for the demonstration and validation of the LAGDA assay.***

Panel Summary

The Panel did not have concerns with the four compounds chosen (prochloraz, 4-*t*-octylphenol, trenbolone, and benzophenone 2). The logic behind the choice of these substances was clear, and each of the compounds is good for proof-of-concept outcomes of the assay. However, because the purpose of these studies was to validate the LADGA assay, additional representative chemicals should be chosen to enhance understanding of a multitude of AOPs.

The Panel had many issues with the statistical methods chosen and described in Section 4.4 of the LAGDA ISR. In general, the statistical methods should be updated, written more clearly and described in detail for each endpoint, and not described as a general set of methods applicable for any endpoint. When combined with the lack of specificity in the statistical methods, this led to incorrect analyses, ranging from treating categorical or ordinal data as continuous to failing to include appropriate random or repeated measures effects in the model. For example, StatCHARRMS, had code that provided incorrect analyses. Overall, the Panel concluded that the statistical methodology is not readily transferable because there appears to be several instances where the protocol was not implemented as planned. Further, there is a serious issue with the failure to identify laboratory methods for some of the endpoints (e.g., the ELISA, T4 and VTG). In addition, the lack of consistency among pathologists argues for a validation exercise or training mechanism that can provide reproducibility among pathology readings by different scientists. See discussion on transferability in the response to charge question 6.

The Panel indicated that the statistical analyses should also include testing across laboratories to identify where differences exist as a means of identifying the more problematic aspects of the testing methods and endpoints. The Panel had a range of specific comments on and recommendations for the

statistical methods. A detailed discussion of all statistical issues and recommendations for all four Tier 2 assays is found in the section entitled, “General Comments Concerning Statistical Issues”.

*Question 5. Considering the variability inherent in biological and chemical test methods, a test method needs to be repeatable and reproducible. A test is robust and reliable if the results are repeatable and reproducible within a laboratory and between different laboratories, respectively. A test protocol should provide sufficient guidance to ensure proper and consistent performance across labs and chemicals. **Please comment on the test method robustness and reliability and the repeatability and reproducibility of the results obtained with the LAGDA assay.***

Panel Summary

Overall, the inter-laboratory testing yielded mixed results. Very few endpoints showed consistent results across laboratories, especially for the thyroid and developmental endpoints and, among the reproductive endpoints, the oviduct was the only endpoint in males and females where a level of consistency was demonstrated. The LAGDA ISR noted that some of the inter-laboratory variability may have been due to technical difficulties with the procedures experienced in some of the laboratories or their failure to abide by test guidelines. This may indicate a problem with the transferability of the assay in its present form. The Panel concluded that reliability and robustness have not been fully demonstrated for the LAGDA assay.

*Question 6. The test protocol should be descriptive enough to be fully transferable to a functional laboratory. The protocol should describe the methodology of the assay in a clear and concise manner so that a laboratory could comprehend the objective, conduct the assay, observe and measure prescribed endpoints, compile and prepare data for statistical analyses, and report the results. Sections 5 and 6 of the LAGDA ISR outline the process of and challenges experienced in the inter-laboratory studies. **Please comment on the transferability across labs and provide any suggestions or recommendations for improvement of the LAGDA assay.***

Panel Summary

The Panel stated that there were significant issues with transferability of the test protocols across laboratories. Only for a few endpoints were there consistent results across participating laboratories. Even the most consistent results in the inter-laboratory analysis, effects on liver and kidney histopathology, exhibited LOEC values that differed by more than 10-fold. Only one endpoint, the delay in oviduct involution in juvenile males, was fairly consistent across all laboratories. Because differences in outcomes were more common than similarities, the Panel concluded that the ability to transfer the technology across laboratories is questionable.

The Panel made the following recommendations to enhance transferability of LAGDA: The protocols need to provide step-by-step details on husbandry, water quality, treatment exposure, tissue collection, quantification of morphological endpoints, histopathology, biochemical assays and statistical analysis. The protocols should be more detailed and specific. In addition, directed training should be provided to any laboratory contracted to run the LAGDA. As stated for all four Tier 2 assay, positive

controls should be included in the LAGDA. Some specific suggestions include the addition of T4 for thyroid endpoints, estradiol or ethynyl estradiol for estrogenic endpoints and trenbolone for androgenic endpoints. Additionally, compounds that act as antagonists to these hormonal systems may also be useful. Having at least one of these compounds used in each run of the LAGDA (chosen based on the potential MOA detected in the Tier 1 assays) will provide quality control to insure the assay is functional in the hands of the contract laboratory.

Question 7. The purpose of the validation process is to determine the readiness of a test for inclusion in a testing program. A component of readiness of a test is the evaluation of the usefulness and limitations of the test, including the classes and types of test substances that can and cannot be tested. Please comment on the strengths and/or limitations of the LAGDA assay.

Panel Summary

In summary, the LAGDA, as a tool for risk assessment of endocrine disrupting chemicals, has the potential to identify the risk of different adverse outcome pathways in *X. laevis* that may be applicable to other wildlife. Unfortunately, high inter-laboratory variation indicates that the validation efforts have not demonstrated the ability to achieve consistent results across laboratories. Without reliable results in the LAGDA making broader extrapolation of the results to wildlife species is difficult.

The Panel had a range of specific comments on and recommendations for the statistical methods (see detailed response to charge question 4). The Panel recommended that clearer and more stringent protocols and QA/QC parameters for performance should be developed and implemented, positive controls should be incorporated into each run of LAGDA and more appropriate application of statistics, including power analysis, should be implemented into the LAGDA protocol before this test may satisfy the stated purpose as a Tier 2 ecotoxicity test.

A detailed discussion of all statistical issues and recommendations for all four Tier 2 assays is found in the section entitled, “General Comments Concerning Statistical Issues.”

Panel Summary of the Mysid Two-generation Toxicity Test (MTTT) and Harpacticoid Copepod Development & Reproduction Test (HCDRT) Charge Questions

Question 1. A rationale for the test method should be available, including a clear statement of scientific basis and the regulatory purpose and need for the test method. The EDSTAC introduced the Tier 2 tests as having the purpose to “characterize the nature, likelihood, and dose-response relationship of endocrine disruption of estrogen, androgen, and thyroid in humans and wildlife.” Although the hormones produced and used by invertebrates are not directly analogous to those of vertebrates (e.g., estrogen, androgen, and thyroid), growth, reproduction, development, and other aspects of invertebrate physiology and life cycle are known to be under endocrine control. EDSTAC went on to note that “chemicals that affect these vertebrate hormones may also affect invertebrate hormones resulting in altered reproduction, development, and growth.” Tier 2 tests were designed to be definitive tests which generate sufficient data to characterize the specific hazard of the substance and

provide sufficient information on dose-response and adverse effects to permit risk decisions. Please comment on the rationale and purpose of the assay as part of the Tier 2 testing in the EDSP, as described in Sections 2.1 and 2.2 of the MTTT ISR.

Panel Summary

The Panel endorsed the concept of the mysid two-generation test as a Tier 2 assay and indicated that this assay is a very sensitive test to measure the possible environmental impacts of toxic or endocrine disruptive compounds on estuarine invertebrates. Invertebrates are significant parts of global ecosystems and thus must be protected from adverse effects of chemicals that disrupt their endocrine systems. Invertebrate endocrine systems will respond quite differently than vertebrate endocrine systems; thus, they need to be addressed in terms of the ecotoxicological risk assessment for endocrine disrupting chemicals. Both the mysid and harpacticoid copepod assays are relevant assays for addressing endocrine disrupting chemical effects in invertebrates.

The Panel was concerned that the use of Tier 2 assays, without a representative invertebrate Tier 1 test or Tier 2 Pre-Screen, will provide less than definitive conclusions regarding a chemical's endocrine disruption potential in invertebrates. The Panel was also concerned that without a Tier I invertebrate assay, there would be no clear rationale for conducting a Tier 2 invertebrate assay. The Panel recommended the EPA consider whether the mysid or copepod multigenerational Tier 2 screens could be effectively modified to develop a Tier 1 screen.

Question 2. Test methods and their associated endpoint(s) should be scientifically relevant to the biological processes of interest and should be demonstrated to be responsive to the specific type of effect/toxicity of interest. Each species presents unique characteristics from a biological perspective and allows for specialized endpoints to address a specific toxicological mode of action. Please comment on the biological and toxicological relevance of the assay in regards to the stated purpose of characterizing endocrine disruptors, as described in Section 2 of the MTTT ISR.

Panel Summary

The Panel stated that both the mysid and copepod Tier 2 assays should be used to quantify interactions with the endocrine system and provide data to predict population-level effects. As such, the current multigenerational mysid and copepod assays fall short in directly quantifying interactions with the endocrine system because there is little known about the endocrine system of mysids, in particular, and invertebrates, in general. Growth and development, which are some of the key test endpoints, are tied to molting, which is controlled by invertebrate hormones. Both protocols do not directly measure molting or the hormone(s) that control this activity. The Panel endorsed the concept of the mysid two-generation test as a Tier 2 assay and indicated that this assay was a very sensitive test to measure the possible environmental impacts of toxic or endocrine disruptive compounds on estuarine invertebrates. The harpacticoid copepod test is an additional bioassay that has been developed to assess endocrine disrupting chemical multigenerational effects in invertebrates. The Panel

recommended the use of the OECD toolbox approach, which includes both the mysid and copepod multigenerational tests that can be used to protect the large diversity of invertebrates.

The Panel made the following recommendations to improve the tests:

- 1) Addition of molecular and metabolomics endpoints may help better define the linkage between growth, development, survival and reproduction with endocrine disrupting chemical related mechanisms of action.
- 2) Future inter-laboratory calibration exercises should include positive controls. Positive controls should be added using methoprene or another known invertebrate endocrine disrupting chemical. Initially, positive control would be tested at multiple doses, then once a dose with a positive response is identified, then only a single dose will be needed for a positive control.
- 3) The EDSP should consider the Hazard Analysis Critical Control Point (HACCP) approach used by other agencies for other regulatory purposes in reviewing the current crustacean endocrine disrupting chemical bioassays. The performance criteria should identify those test endpoints deviations that represent major and minor non-conformities.

*Question 3. The test protocol should be sufficiently detailed and should include a description of what is measured and how it is measured. The selection of endpoints within the assay should be reflective of the biological processes of interest and the endpoints should be intrinsically relevant and have established sensitivity. The test protocol should demonstrate the ability to measure the endpoints and provide adequate performance criteria for evaluation. **Please comment on the selection, optimization and demonstration of the assay endpoints, as outlined in Sections 3 and 4 of the MTTT ISR.***

Panel Summary

The selected endpoints are appropriate for detecting the apical adverse effects of regulatory concern that vertebrate EAT AOP active chemicals may have on invertebrates, regardless of whether or not they disrupt the crustacean endocrine system. The optimization and demonstration of the endpoints appears sufficient. However, the amount of inter-laboratory variability and general disagreement in which endpoints are affected and at what dose, suggests that these tests are not yet reliable and repeatable. This suggests that further optimization and demonstration of the mysid and copepod assays is needed.

*Question 4.1. Demonstration of the test method performance should be based on the testing of reference chemicals representative of the types of substances for which the test method will be used. Test substances should adequately represent an appropriate range of responses and physical/chemical properties for which the test method is proposed to be appropriate. The selection of the most appropriate statistical approaches depends in part on the nature of the data and also on the design of the validation study. Statistical and non-statistical methods used to analyze should be described. **Please comment on the selection of test substances and methods (analytical and statistical where appropriate) chosen for the demonstration and validation of the MTTT assay.***

Panel Summary

The Panel recommended that the choice of experimental design and statistical methods used in demonstration and validation of the MTTT assay be described in greater detail. This recommendation is common to all four Tier 2 ecotoxicity tests reviewed by the Panel. The Panel reviewed Battelle's re-analysis of the data from the various laboratories because it had the most consistent and clearest description of the methodology used to analyze the inter-laboratory data. The Panel had many comments and recommendations concerning the MTTT statistical analysis (see detailed response for this charge question).

Question 4.2. The proposed MTTT study design uses three replicates per control and treatment group and additional replicates require greater resources. Please comment on the proposed level of replication and the subsequent statistical analysis of the data.

Panel Summary

The Panel recommended that a threshold number of replicates are needed, i.e., minimal value of replicates, to continue with the F1 exposures regime. One Panel member suggested that the number of replicates might be increased by reducing the number of individuals within each tank and adding more tanks (which will not add more total individuals). There is an interrelationship between numbers of replicates and numbers of organisms per replicate and the combined influence on statistical power. This relationship will vary with species and test design (among other factors). The Panel suggested some further analysis using existing data to generate random generated data sets with larger replicate size to examine the effect of sample size on power so that sample size can be optimized. The Panel emphasized the importance of developing a statistical manual for the statistical analysis methods, approaches and procedures to follow when conducting all four Tier 2 ecotoxicity tests, including the MTTT.

Question 5. Considering the variability inherent in biological and chemical test methods, a test method needs to be repeatable and reproducible. A test is robust and reliable if the results are repeatable and reproducible within a laboratory and between different laboratories, respectively. A test protocol should provide sufficient guidance to ensure proper and consistent performance across labs and chemicals. Please comment on the test method robustness and reliability and the repeatability and reproducibility of the results obtained with the MTTT assay.

Panel Summary

The Panel indicated that two laboratories that completed the study had results that led to different conclusions. The Panel suggested that this may be due to: 1) the sample sizes are simply too small for reproducible analyses, 2) no treatment effects on the measured endpoints and all significant results were type I errors, 3) the appropriate concentrations of chemicals were not used to show effects, or 4) the laboratories failing to observe and follow the protocols as directed. The Panel could not determine whether the statistical methods were at fault since there was insufficient information to determine whether the tests for checking assumptions were appropriate or adequate and how the decision for

three replicates was made. Sample sizes of three are much too small to conduct most of the statistical tests being applied in the subsequent analyses. The power analysis was not informative.

The Panel concluded that the test method is not robust or reliable and the results are not reproducible or repeatable. The Panel agreed with EPA's recommendations to increase the feeding regiment, number of F0 animals at the onset of the assay and removal of the self-starting siphons. An evolution of some endpoints occurred over the course of the studies which allowed for more concise measurement, but disallowed direct comparisons between tests. A detailed discussion of all statistical issues and recommendations for all four Tier 2 assays is found in the section entitled, "General Comments Concerning Statistical Issues".

The Panel agreed that pairs of surviving mysids that do not produce a second brood should not be included in this count, but rather should be included in the "% females reproducing" endpoint. However, since this is not a sensitive endpoint and the numbers of mating pairs were very low in many of the studies, the Panel recommended that the EPA assess the loss of power in the second endpoint.

*Question 6.1. The test protocol should be descriptive enough to be fully transferable to a functional laboratory. The protocol should describe the methodology of the assay in a clear and concise manner so that a laboratory could comprehend the objective, conduct the assay, observe and measure prescribed endpoints, compile and prepare data for statistical analyses, and report the results. Section 5 of the MTTT ISR outlines the process of and challenges experienced in the inter-laboratory studies. **Please comment on the transferability across labs and provide any suggestions or recommendations for improvement of the MTTT assay.***

Panel Summary

The Panel indicated that both assays were not yet transferable across laboratories. There were issues of poor compliance with all test protocols and some inconsistencies in the data for growth, development and reproductive endpoints among different laboratories conducting the mysid test. The copepod test was not a familiar test for the testing laboratories, where there were problems achieving test concentrations in the small dosing chambers of the 96-well plate and the laboratories were less familiar with the life history stages. The Panel envisioned these issues could all be corrected by modifying the current test protocols and by additional training. One major issue for both tests is lack of information concerning the invertebrate endocrine response, given that most test endpoints are indirect assessment endpoints for invertebrate endocrine disrupting chemicals.

The Panel identified three areas that need to be addressed to improve the transferability of the current MTTT bioassays.

- 1) Use of positive controls.** The Panel recommended that future inter-laboratory calibration exercises include use of positive controls with known invertebrate endocrine disrupting chemical effects and be run at concentrations that should always elicit these responses. For example, methoprene or other known invertebrate endocrine disrupting chemicals would be

run initially at multiples doses, but after many runs and confidence increases in the results, only a single dose of the positive control would be needed for each new compound tested.

- 2) **Importance of defining major/minor deviations in the laboratories performing the assay.** The EPA would greatly benefit from a Hazard Analysis Critical Control Point (HACCP) approach used by other agencies for other regulatory purposes in reviewing the current crustacean endocrine disrupting chemical bioassays. EPA should identify those test endpoints deviations which represent major and minor non-conformities in the protocol.
- 3) **Experimental design and statistical methods.** The Panel stated there were several issues concerning the description of the experimental design and statistical methods in the MTTT ISR.

The Panel agreed with the proposal made by the EPA (see section 4.1.3 of the MTTT ISR) to improve the protocol, but added the following recommendations: increase the number of organisms used to initiate the test, provide training in how to run the assay and add a positive control that is a known invertebrate endocrine disrupting chemical, e.g., methoprene. Sample sizes of three are much too small to conduct most of the statistical tests being applied in the subsequent analyses.

Question 6.2. Based on the validation results using the two different invertebrate species, mysid (MTTT) and copepod (HCDRT), the EPA is proposing the mysid protocol as the preferred Tier 2 assay. Please comment on the rationale to recommend the mysid protocol as the preferred Tier 2 invertebrate assay, as described in Section 6.4 of the MTTT ISR.

Panel Summary

The Panel considered the choice of the mysid over the copepod assay at this point to be premature based on the results presented to date. Both assays have similar chemical sensitivities even at low concentrations. The Panel recommended inclusion of both the mysid and copepod bioassays rather than selecting one bioassay. The Panel agreed with the OECD recommendation that both bioassays (classified by OECD as Level 5 tests) be used to provide additional data to enhance invertebrate protection. The OECD toolbox approach for testing invertebrates recognizes the broad and diverse biology of invertebrate species (especially important are the crustaceans, mollusk, annelids, etc.).

Question 7. The purpose of the validation process is to determine the readiness of a test for inclusion in a testing program. A component of readiness of a test is the evaluation of the usefulness and limitations of the test, including the classes and types of test substances that can and cannot be tested. Please comment on the strengths and/or limitations of the assay, as described in Section 6 of the MTTT ISR.

Panel Summary

The Panel summarized the strengths and limitations of the MTTT, and provided recommendations to address the limitations and strengthen the assay's performance (see detailed response to this charge question).

The MTTT has many advantages including familiarity and extensive experience of testing laboratories with this organism. Some of the advantages of this assay include:

- 1) A variety of test endpoints are used to assess alterations in growth, development and reproduction which are integrated to predict endocrine disrupting chemical effects.
- 2) The test endpoints are generally simple and easy to measure.
- 3) The assay includes endpoints that are relevant to population-level effects and the results are amenable to common population modeling approaches.
- 4) Males and females are distinguishable by conventional microscopy based on anatomical differences.
- 5) The results that showed F0 > F1 for certain endpoints were more, or as sensitive, as the results for the F2 endpoints. This underscores the importance of the current mysid testing used by EPA for pesticide registration as an important source of information.
- 6) The addition of the second generation assessment for the mysid test indicates the importance of additional maternal transfer exposure in assessing invertebrate endocrine disrupting chemical effects.
- 7) Test is recommended as part of the OECD testing protocol (Level 5).

The Panel also described the limitation of the MTTT, two of the major limitations are:

- 1) The large number of endpoints included in the test makes the test tedious. The apparent redundancy among endpoints suggests that they may not have been selected strategically.
- 2) The variability among laboratories in the ring test is troublesome. This variability includes inability of some laboratories to achieve control performance metrics suggesting that testing laboratories are not proficient in conducting this test and/or that the test endpoints are highly variable. Earlier studies with mysids and other invertebrate species have indicated it is not possible to predict life sensitivity in the F1 generation. Similarly, the MTTT cannot predict generational sensitivity among life stage endpoints. This variability also includes endpoints between replicates, treatments, and laboratories. Whether this variability is intrinsic (e.g., animals, feeding, feed quality) or is due to test design or lack of laboratory proficiency is uncertain, but it makes interpretation of the existing data speculative at best. Further, given this variability, it is uncertain if laboratories repeating the bioassay would get the same results.

The Panel agreed with the EPA's recommended changes to the MTTT protocol and added the following three recommendations that will more completely address all the limitations in the protocol

- 1) Increase the number of organisms used to initiate the test to increase probability of survival and adequate reproduction.
- 2) Conduct additional performance studies to increase confidence that each laboratory can conduct all assays according to established protocols and under GLP.
- 3) Add a positive control that is a known invertebrate endocrine disrupting chemicals, e.g. methoprene, with a described invertebrate hormonal pathway effect for inclusion in future testing. A positive control will help establish growth, development and reproductive “fingerprints” that may be useful in better discerning invertebrate endocrine disrupting chemical effects in screening new compounds.

Question 8. There is sufficient evidence to indicate that endocrine disrupting chemicals can disrupt normal development and reproductive success, however the sensitivity of the F1 generation compared to the F0 is less clearly defined. Please comment from a scientific and risk assessment perspective on the value added of multiple generations in the MTTT assay.

Panel Summary

The use of multiple generations is intended to capture effects that may not be apparent through exposure of a single generation. As demonstrated by comparison of the mysid sensitivity to chemical exposure by generation, LOECs for the F1 generation can be more sensitive than LOECs observed in the F0 generation. The addition of multiple generations to the mysid test provides information pertinent to risk assessment that would be lost if the F1 generation was removed from the test. In some studies, it was apparent that the number of organisms carried into the F1 generation tests was not sufficient to generate statistically meaningful results especially after loss of animals (see the response to charge question 5 for more details). As a result, the Panel recommended that the F1 generation study be continued; however, the protocol should be amended to use larger sample sizes in the F0 generation so that a larger number of animals are available for continuation of tests in the F1 generation study.

DETAILED DISCUSSION OF THE CHARGE QUESTIONS

Japanese Quail Two-Generation Test (JQTT) Charge Questions

*Question 1. A rationale for the test method should be available, including a clear statement of scientific basis and the regulatory purpose and need for the test method. The EDSTAC described the Tier 2 tests as having the purpose to “characterize the nature, likelihood, and dose-response relationship of endocrine disruption of estrogen, androgen, and thyroid in humans and wildlife.” Tier 2 tests were designed to be definitive tests that generate sufficient data to characterize the specific hazard of the substance and provide sufficient information on dose-response and adverse effects to permit risk decisions. **Please comment on the rationale and purpose of the assay as part of the Tier 2 testing in the EDSP, as described in Sections 2.5 and 2.6 of the JQTT ISR.***

Panel Response

The Panel concluded that the JQTT is scientifically justified and clearly has great potential as a standardized multi-generation reproduction assay for use in the assessment of endocrine disruptor effects on birds; however, the test protocol in its current form needs some refinement, particularly in terms of the endpoints measured, guidance for transferability across labs and data analysis. The JQTT has a great deal of merit in terms of measurement of the effects of endocrine disrupting chemicals in Galliform birds (Order Galliformes), but it is not clear how broadly applicable the findings will be to other orders of birds with very different reproductive life-histories. For example, phallus size in many ducks is an important reproductive character that is lacking in Galliformes. Perhaps more importantly, in the highly diverse Passerine birds (order Passeriformes), song control system differentiation and size, song behavior, and effects on female choice and sexual behavior could all be affected by endocrine disrupting chemicals used in an agroecosystem, such as the Great Plains populated by passerines, and would not be detected by using Japanese quail in this test. Given this situation, the Panel strongly recommended that a multi-generation avian reproduction test be developed on an altricial, passerine bird, e.g., zebra finch (*Taeniopygia guttata*), to complement the JQTT. The Panel provided additional recommendations concerning immune endpoints.

1) Purpose of the test

Overall, the Panel agreed that the EPA (section 2, JQTT ISR and appendices) provided detailed information to support the scientific rationale and purpose of an avian two generation test using Japanese quail as a representative bird (JQTT).

The science is clear that birds may react in fundamentally different ways than mammals to endocrine disrupting chemicals, e.g., differences in the mechanism of sex determination with males as the homogametic and females the heterogametic sex, differences in specificity of hormone receptor binding and other aspects of hormonal axes among distantly related taxa (summarized in section 2.5 of the JQTT ISR). Given the EDSP goal of using Tier 2 ecotoxicity tests in risk assessments, it is important

to have taxon-specific information for birds. In both aquatic and terrestrial ecosystems, birds will be an important focus of possible ecological risk assessments. Information from the Tier 2 tests will be critical for evaluating risks to threatened and endangered species, as well as other birds of conservation concern, migratory birds protected by federal law, and game birds managed by both state and federal agencies. The need for the JQTT goes beyond assessments focused on birds. Evaluation of endocrine disruptor effects on wildlife is included in EDSP's mandate. Under the currently proposed tests, the JQTT is the only protocol specifically designed to evaluate endocrine disruptor effects on terrestrial wildlife. The laboratory rodent (mammalian) tests are first primarily intended to inform human health assessments. This rationale further emphasizes the regulatory purpose and need for the JQTT.

2) Need for the test method

While other avian toxicity test protocols are already in use (e.g., avian dietary toxicity test, OCSPP 850.2200), there is a need for a multi-generation avian reproduction test method to look specifically at endocrine disruptor-relevant endpoints. The Agency provided sufficient rationale in the JQTT ISR and supporting appendices, e.g., OECD JQTT guidance in Appendix B, to document why the JQTT is needed as a Tier 2 ecotoxicity test. Specifically, these documents outline the arguments to support the need for additional endocrine-relevant endpoints important to survival and hence a population. The Panel also indicated a need for inclusion of additional endpoints, which are clearly under endocrine regulation and are also directly relevant to population endpoints. The protocol optimally should include endpoints such as female and parental reproductive behavior. These are not seen in Japanese quail and were part of the rationale for investigating zebra finch as a second avian species for testing. Moreover, there is a need for the addition of endpoints of the immune system functioning in the JQTT in view of the interactions of steroid and other hormones on immune functioning and the importance of the immune system to individual animals and to populations.

Given the Panel's understanding that effects of some endocrine disruptor chemicals depend on what stage of development a bird is exposed, the protocol needs to consider all stages of the life cycle. Finally, research results are increasingly indicating that endocrine disruptor chemicals can have cross-generation effects supporting the need for a multi-generation test (addressed in more detail in response to charge question 8). The protocols approved by the American Society for Testing and Materials (ASTM 1990), OECD (1993) and EPA (2012 a, b) lack these important components, which supports the Panel's analysis of the need for the JQTT test.

3) JQTT as a definitive test

The charge question states that the *“Tier 2 tests were designed to be definitive tests that generate sufficient data to characterize the specific hazard of the substance and provide sufficient information on dose-response and adverse effects to permit risk decision.”* The Panel agreed that Japanese quail is an appropriate test species for this protocol; however, there are limitations on how well it can fully characterize adverse effects of endocrine disrupting chemicals on the range of birds that might be covered in some environmental risk assessments. In this sense, the Panel concluded that the JQTT would not be viewed as a “definitive test.”

The OECD report (Appendix D, JQTT ISR) summarizes the rationale for choosing Japanese quail over Bobwhite quail for the avian two generation reproduction test. This report clearly and convincingly outlines why Japanese quail are a more appropriate test species than Bobwhite quail. However, the summary of the advantages of Japanese quail in the OECD report also clearly shows that the characteristics in favor of this species are strictly practical in nature; i.e., they are domesticated species, easy to raise in captivity, have an unusually short generation time, and their precocial development allows for production of large numbers of eggs that can be artificially incubated to produce the next generation.

The Japanese quail offers a series of specific advantages. These are summarized in the ISR and discussed in detail in the Panel's response to charge question 7. In particular, Japanese quail have a short generation time and are readily reared in the laboratory. In general, they can be used as a surrogate for wild birds, but there are limitations. The Panel stressed the importance of ensuring that this test protocol is logistically practical to implement if these tests are to be successfully used on a large scale for regulatory testing.

The unique features that make Japanese quail a good laboratory subject, also mean it is not representative of most wild birds (see limitations discussed in the Panel's response to charge question 7). An important limitation is the representativeness of Japanese quail to a broad range of wild bird species that might be included in an environmental risk assessment. Japanese quail and their relatives (Order: Galliformes, Family: Phasiandidae) are found in a phylogenetic lineage (super-order Galloanserae) that diverged from the vast majority of birds (Superorder: Neoaves) around 100 million years ago (see Tree of Life project website: <http://tolweb.org>) and also Wang *et al.* 2011, Hackett *et al.*, 2008, Jetz *et al.*, 2012, McCormack *et al.*, 2013, etc.). Jones *et al.* (2013) point out the limitations of depending on Japanese quail in the test and emphasize they are not representative of all birds. While many characteristics of the endocrine system will be conserved, this degree of divergence undoubtedly leads to some differences including vulnerability to endocrine disrupting chemicals.

Japanese quail exhibit precocial development (well developed at hatch) while most birds are altricial (poorly developed at hatch), Japanese quail do not have the array of female sexual and parental behaviors found in many avian species. Japanese quail are relatively short-lived. It is therefore difficult to extrapolate to long-lived species. The difficulty in extrapolating to other avian species is not limited to long-lived species as noted in the JQTT ISR.

Extrapolation across bird species may occur at several different levels (Perkins & Garcia-Reyero, 2013). The simplest is at the molecular initiating event level where the sequences or structures of specific proteins critical to initiating chemical effects can be compared. For example, the sequence of receptors or enzymes can help predict the likelihood that a similar chemical-target interaction would occur across bird species. Identification of a protein ortholog for a known chemical molecular target can be used to infer possible effects, especially where an adverse outcome pathway exists. Numerous human drug targets are conserved across ecologically-relevant vertebrate and non-vertebrate species (Gunnarsson *et al.*, 2008). However, assigning functions via sequence similarity should be done with

caution as genome duplication in some species, such as zebra finch, has resulted in multiple orthologs for 15% of human genes (Howe *et al.*, 2013).

More complex extrapolations occur at a pathway level, from molecular initiating event to adverse outcome, including the sequence of events and the dose or threshold concentrations required to activate these events. Pathways can be explored either as discrete pathways or as networks using cross species comparative genomics (for a brief review see Burgess-Herbert & Euling, 2011). Ultimately, the most realistic approach is to translate potential species specific effects through systems level models where dynamic events are incorporated, such as chemical concentrations, homeostasis, measuring effects over time, and species-specific parameters related to absorption, distribution, metabolism, and excretion. This may be especially important where physiology of the species is different, such as a more rapid metabolism or heartbeat. If target pathways and receptors are highly conserved between zebra finch and Japanese quail, for example, then it would be possible to predict that the two animals would behave similarly, assuming similar toxicokinetics and toxicodynamics.

In summary, while the data produced using the JQTT will provide a valuable and necessary contribution to inform risk decisions, the Panel did not believe that the JQTT is the “definitive” test that would “provide sufficient information” to conduct an ecological risk assessment for wild birds. The limitations inherent in relying on Japanese quail described here, plus the lack of information about altricial species and many ecologically important endpoints (e.g., parental care, female mating behavior and immune function) support the need for additional data from other bird species and perhaps other tests.

4) Immune function. The Panel indicated that androgenic or anti-androgenic endocrine disrupting chemicals would influence avian immune functioning. The Panel suggested that other endpoints such as proliferations of lymphocytes and primary antibody responses to a challenge be added to the protocol.

Recommendations

- 1) The Panel recommended that the limitations of the JQTT be carefully outlined and guidance be drafted to describe situations when additional information about other groups of birds may be needed to better characterize the risk (i.e., lower the uncertainty in extrapolating the results of the JQTT) and better inform risk management decisions.
- 2) The Panel strongly recommended incorporation of a second species, zebra finch, as a complementary multi-generation test to the Japanese quail. As noted above, the JQTT is the only proposed Tier 2 tests specifically intended to evaluate the effects of endocrine disruptor chemicals on terrestrial wildlife. As such, the Japanese quail results will be very broadly extrapolated to a diverse array of taxa. The Panel indicated that there may be environmental situations in which Japanese quail may be not representative enough to extrapolate results to some wild bird species. The gap between the proposed tests that rely only on the Japanese quail and the goal of being a “definitive test” that “fully characterizes adverse effects” could be partly addressed by incorporating supplemental data from an altricial passerine. Addition of a second

species would provide a good balance between ease of use in the laboratory and ability to generalize the results to wild species when specific bird populations are included in the use area of a potential endocrine disrupting chemical (Jones *et al.* 2013). This is the main rationale why a complementary Tier 2 test using zebra finches or other passerines will be valuable.

The zebra finch has the advantage of being a passerine (Order: Passeriformes) species with altricial development, extensive female sexual and parental behavior and a published genome. The order Passeriformes includes more than half of all bird species. While there are drawbacks to the zebra finch relative to other passerines, the Panel recommended its use because it is a model laboratory organism with many of the same practical advantages associated with the choice of Japanese quail. Zebra finches breed readily in captivity and, like Japanese quail, have a relatively short generation time. Zebra finch hatchlings reach sexual maturity in approximately 90 days (Zann, 1996). This timeline would allow a zebra finch protocol to be completed in approximately 20 weeks. The ability to raise zebra finches successfully in the laboratory has led to its use as a model organism. Research using zebra finches has provided valuable information about overall development, behavior and functioning of the nervous and endocrine systems that will facilitate interpretation of test results. Rapid advances in our understanding of zebra finch biology, especially the recent publication of the zebra finch genome (Warren *et al.*, 2010) will facilitate future advances in testing.

The general test design presented for the JQTT could be easily adapted to the zebra finch by incorporating the recommended modifications made by the Panel throughout this report. The Panel recognized that it will not be beneficial to run duplicate protocols with both Japanese quail and zebra finch for all chemicals. The Panel recommended that the test protocol include explicit descriptions of the limitations of the JQTT for fully characterizing possible adverse effects and guidance for when additional information from other groups of birds will be needed for ecological risk assessments. Existing information about endocrine pathways and relevant endpoints could be synthesized with the limitations of the Japanese quail as a model species. The purpose of this information would be to explain the circumstances when the Japanese quail is not a sufficient model for detecting possible adverse effects. For example, existing work shows that male sexual behavior in both altricial and precocial species is profoundly impacted by exposure to 17 β estradiol during development. The vulnerable period in precocial species occurs prior to hatching so maternal deposition is a key exposure window, while the vulnerable period in altricial species occurs post-hatching. Feeding during the first week post hatch is the key exposure window for altricial species. Likewise additional information would be needed in cases where key reproductive behavior endpoints (e.g., song, female choice, parental care) are vulnerable. In addition, cognitive abilities (learning, memory, spatial ability) have direct links to survival in wild passerines and are well studied in zebra finches. When existing knowledge or the results of other Tier 1 or Tier 2 tests indicate that these or similar key endpoints or pathways are vulnerable to a chemical undergoing testing, then the Japanese quail results should be supplemented with additional data such as zebra finch tests.

- 3) Emerging advances in avian genomics might be employed to evaluate when the mode of action of a given chemical is likely to differ among bird species. The simplest approach is to focus on the molecular initiating event level where the sequences or structures of specific proteins critical to initiating chemical effects can be compared. Pathways can be explored either as discrete pathways or as networks using cross species comparative genomics. A realistic approach is to examine potential species specific effects through systems level models where dynamic events are incorporated, such as chemical concentrations, homeostasis, measuring effects over time, and species-specific parameters related to absorption, distribution, metabolism, and excretion.
- 4) The Panel recommended that immune endpoints be added to the JQTT. At autopsy, weight of the bursa *Fabricius*, thymus and spleen should be determined. The differential count of leukocytes should also be determined. There is abundant evidence of the sensitivity of the differential leukocyte counts to toxicants. The bursa of *Fabricius* (Latin: *Bursa cloacalis* or *Bursa fabricii*) has a critical role in B cell development and humoral immunity in birds (Chang *et al.*, 1955; Glick *et al.*, 1956) and in the ability of androgens to completely suppress bursal development (Meyer *et al.* 1959, Mueller *et al.* 1960; Bruggeman *et al.* 2003). The Panel strongly encouraged EPA to conduct research on effects of endocrine disrupting chemicals on immune functioning in the JQTT. This would allow the development of immune endpoints. See also response to charge question 1.
- 5) The Panel recommended that EPA conduct research to study epigenetic effects and once identified, evaluate the importance of specific epigenetic endpoints in endocrine disruption. The addition of epigenetic endpoints is discussed in detail in the response to charge question 7.

Question 2. Test methods and their associated endpoint(s) should be scientifically relevant to the biological processes of interest and should be demonstrated to be responsive to the specific type of effect/toxicity of interest. Each species presents unique characteristics from a biological perspective and allows for specialized endpoints to address a specific toxicological mode of action. Please comment on the biological and toxicological relevance of the assay in regards to the stated purpose of characterizing endocrine disruptors, as described in Sections 3 and 4 of the JQTT ISR.

Panel Response

The Panel concluded that the test methods and their associated endpoints for the JQTT are scientifically relevant to the biological processes of interest. The Panel suggested a number of improvements to the protocol that could be incorporated immediately. In addition to the protocol improvements, the Panel strongly encouraged research in the following areas: 1) zebra finch as an additional representative species to complement the Japanese quail and 2), effects of endocrine disrupting chemicals on immune functioning, and potential epigenetic effects (see also responses to charge questions 1 and 7). A detailed discussion of the endpoints is found in the Panel's response to charge question 3.

The Panel recommended the following improvements to the protocol, which should increase the repeatability within and between laboratories.

- 1) Use sixteen pairs of Japanese quail to ensure that at least 12 pairs remain for completion of generation F0 studies;
- 2) Add specifications for water quality;
- 3) Add a comprehensive list of diet requirements with appropriate references; the nutritional requirements are presented in Table 3-2 (JQTT ISR);
- 4) Increase the number of replicates for hormone assays and histopathology, e.g., at least eight replicates to compensate for lost samples;
- 5) Maintain the eggshell endpoint because environmental toxicants influence reproductive success via defects in eggshells with one of the most widely cited examples being the reduction in shell thickness with DDT (Hickey and Anderson, 1968; Bitman *et al.*, 1969; Davison *et al.*, 1972; Kolaja and Hinton, 1977; Holm *et al.*, 2006); and,
- 6) Maintain inclusion of behavioral endpoints because birds may display unique behavioral responses caused by exposure to potential endocrine disruptors.

After substantial debate, the Panel concluded that there should be flexibility on the dose-response curve depending on the chemical in question.

Question 3.1. The test protocol should be sufficiently detailed and should include a description of what is measured and how it is measured. The selection of endpoints within the assay should be reflective of the biological processes of interest and the endpoints should be intrinsically relevant and have established sensitivity. The test protocol should demonstrate the ability to measure the endpoints and provide adequate performance criteria for evaluation. Please comment on the selection, optimization and demonstration of the assay endpoints, as outlined in Sections 4 and 5 of the JQTT ISR.

Panel Response

The Panel concluded that all of the selected endpoints are scientifically relevant to the biological processes of interest. The Panel concluded that, in general, the JQTT protocol is sufficiently detailed in describing the endpoints measured and how they were measured with some caveats described in the recommendations regarding the selected endpoints. The Panel also recommended additional endpoints should be considered.

Recommendations

- 1) **Growth, development, and reproduction.** The Panel recommended that the JQTT continue to include the following endpoints related to growth, development and reproduction: egg production, growth, embryonic development, sex reversal and mortalities. The endpoints of egg production, growth, embryo development and phenotypic sex (versus genotypic sex) are strong and clearly fully justified for testing endocrine disrupting chemicals for “developmental window” type endocrine disruption and/or apical effects. Strong consideration should be given

to extending the JQTT for a complete F2 egg-laying cycle. Examination of F2 egg laying and male reproductive systems is important. One study on trenbolone (17 β -Hydroxyestra-4,9,11-trien-3-one) went beyond the protocol and included a full assessment of gross reproductive endpoints. This study provided definitive evidence that trenbolone received in the yolk from the dam acted at the embryonic stage on a “window” for reproductive development.

- 2) **Sexual maturation as an endpoint of endocrine disruption.** The data in the trenbolone and vinclozolin studies did not show effects on sexual maturation contrary to the conclusion of the OECD that Japanese quail “*age at sexual maturity was among the most sensitive measures of impaired reproduction in the female Japanese quail exposed to dietary lead from hatch through 12 weeks of age,*” The Panel concluded that this endpoint did not add much value and could be dropped from the protocol.
- 3) **Egg shell thickness.** The Panel recommended that egg shell thickness for F0, F1 and F2 continue to be endpoints for the JQTT as it meets the criteria of reliability and relevance. Moreover, the effect of endocrine disrupting chemical on egg shell thickness is a critical endpoint given the literature on toxicants and endocrine disrupting chemicals such as DDT (Hickey and Anderson, 1968; Bitman *et al.*, 1969; Davison *et al.*, 1972; Kolaja and Hinton, 1977; Holm *et al.*, 2006). The CV for shell thickness between laboratories was very acceptable at 8%. Moreover, there was close agreement between shell thickness for controls in F0 and F1 generations, respectively 0.21 and 0.21. There were, however, marked inter-generational differences in shell thickness in control birds in one laboratory. Standards for intra- and inter-laboratory comparison should be added to the protocol.
- 4) **Behavioral indices.** The Panel recommended that behavioral indices be continued as important endpoints in the JQTT. This series of endpoints meets the criteria of reliability and relevance. Although these endpoints are labor intensive, they add a novel and very important biological parameter. For example, there are marked effects of trenbolone including decreased mounting attempts and the number of successful copulations.
- 5) **Spermatozoan endpoints.** The Panel recommended that spermatozoa characteristics be included as an endpoint for the JQTT as a decline in spermatozoa number or motility are very likely to influence avian populations. Epididymal spermatozoa characteristics exhibited a lack of consistency as is exemplified by the data reported for F0 and F1 in the endosulfan study. The Panel encouraged EPA to conduct further research to adapt existing semen collection techniques from turkeys, chickens and some wild birds in the laboratory to the Japanese quail and determine the applicability of spermatozoan characteristics as an endpoint in the JQTT.
- 6) **Cloacal area.** The Panel recommended that the measurement of the cloacal area is a very good endpoint for endocrine disrupting chemicals with androgenic or anti-androgenic activities. However, the androgen, trenbolone, failed to have a consistent effect on cloacal area. It is not clear whether this parameter is useful and should continue to be part of the JQTT protocol.

- 7) **Hormone assays.** The Panel stated that the endpoints chosen by the EPA are broadly appropriate, useful and meet the criteria of reliability and relevance. The CV between different studies is satisfactory. Specific recommendations are the following:
- Plasma hormonal endpoints should be measured in adult F0 and F1 birds and embryos. In principle, the ability of putative endocrine disrupting chemicals to modify endocrine endpoints in the JQTT provides excellent evidence for mechanism of action and, in particular, effects on a “developmental window” sensitive to endocrine disrupting chemicals. There were not consistent changes in plasma concentrations of hormones in the F2 generation. Consideration should be given to discontinuing plasma hormonal endpoints in F2 birds.
 - Yolk and fecal/urine hormonal endpoints should be discontinued as there was a lack of consistency in responses.
 - The Panel encouraged the EPA to conduct further research on the plasma hormonal endpoints focusing on the adrenal glucocorticoid hormone, corticosterone. Ultimately, the hormonal endpoints in the JQTT should be expanded to include the adrenal stress hormone, corticosterone. This is consistent with the OECD protocol described in Appendix B, JQTT ISR.
 - The numbers of replicates for the plasma hormonal endpoints should be increased to provide sufficient statistical power.
 - Both estradiol and testosterone vary markedly during the daily ovulation cycle. With daily production of eggs, these cycles will be entrained to time of day. Therefore, the Panel recommended that blood sampling should be at a fixed time of day and this instruction should be reflected in the protocol.
- 8) **VTG.** The Panel indicated that the VTG endpoint is not ready for inclusion in the JQTT due to lack of data in the JQTT validation studies measuring the effects of endocrine disrupting chemicals on plasma concentrations of VTG during the validation studies of the JQTT. Therefore, the Panel recommended further research on the effect of endocrine disrupting chemicals on plasma concentrations of VTG and/or hepatic expression of VTG. This endpoint(s) could become a powerful endpoint for measuring estrogenic or anti-estrogenic effects of endocrine disrupting chemicals in birds given the results from previous studies on the estrogenic control of expression/production of VTG in the Japanese quail (e.g., Gupta and Kanungo 1996; Shibuya *et al.* 2005).
- 9) **Epigenetic endpoints.** Epigenetics is “*the study of heritable changes in gene expression that are not due to changes in DNA sequence*” (Eccleston *et al.*, 2007). Examples of the mechanisms underlying epigenetics include methylation of DNA and modification of histones. Epigenetic changes play a critical role in both development and responses to environment stressors or nutritional challenges (Eccleston *et al.*, 2007; Jablonka & Raz, 2009; Ho & Burggren, 2010; Bhandari *et al.*, 2012; Manikkam *et al.*, 2012). At least some epigenetic changes are inheritable, i.e., transmittable to the next and even subsequent generations (Jablonka & Raz, 2009; Ho & Burggren, 2010; Manikkam *et al.*, 2012). The possible organizational changes that might occur during embryonic development and the possible

epigenetic effects are part of what make endocrine disrupting chemicals of special interest. The Panel recommended the Agency pursue research on possible epigenetic endpoints. Perhaps, there are epigenetic factors that are predictive of reproductive and/or other endocrine modifications that might be worth pursuing.

- 10) Additional endpoints measured.** The following endpoints were not measured: GnRH-I and -II, gonadotropin inhibitory hormone (GnIH), luteinizing hormone (LH), follicle-stimulating hormone (FSH) and progesterone. GnRH-II is critical for sexual behavior and can influence Hypothalamic–pituitary–gonadal (HPG) axis activity. LH and FSH are critical components of HPG axis. Progesterone is critical for ovulation and oviduct function. Measurement of these endpoints is important to the investigation of population-level effects.

Many on the Panel recommended that endpoint measurements of GnRH-I and -II (I and II) together with gonadotropin inhibitory hormone (GnIH) in the brain, LH, FSH and progesterone in the blood; luteinizing hormone receptor (LHR), follicle-stimulating hormone receptor (FSHR), StAR protein and 17 β -HSD in the gonads, using qPCR, be included in the protocol. This could be done easily in males on one gonad and then histology could be performed on the other. It could be performed on individual follicles and parts of the oviduct in females (in a standardized manner). Thus, these additional endpoints could be measured without increasing the number of animals used. In the brain GnRH-I and -II can be measured via qPCR and histology (immunohistochemistry or *in situ* hybridization) using alternate brain sections within individuals. More informative data can be gathered without increasing the number of animals needed. Other Panel members did not support including all of these endpoints. Each one would require a great deal of protocol development, assay validation, etc. There is insufficient information available to describe how their inclusion would strengthen specific conclusions about mode of action. Another Panel member considered the inclusion of peptide/protein hormones and neuropeptides at least as representative of the physiological mechanisms underlying reproduction as the gonadal steroids.

Question 3.2. Pathological evaluation of histological tissue preparations is an established, sensitive and integral endpoint in the assessment of effects in long term in vivo assays. The tissues targeted for histopathological evaluation should be shown to be sensitive to exposure and relevant to a mode of action or pathway determination. Tissue samples from several organs or glands (e.g., kidney, liver, thyroid, gonads) underwent histopathological examination in the JQTT inter-laboratory validation and a discussion of histopathology as an endpoint is provided in Section 4.1.13 of the JQTT ISR. Please comment on the value of histopathological analyses in the JQTT assay for each of the tissue types examined and what (if any) critical information is gained from their inclusion or would be lost if histology were not examined.

Panel Response

The Panel concluded that histopathological analyses for all tissue types was well-justified and should be included in the JQTT, i.e., gonadal, epididymal, liver, kidney, and thyroid histopathology. Abnormal histopathology is evidence for endocrine disruption and/or apical effects. The Panel strongly

recommended the use of morphometric analysis as a means to enhance analytical power and consistency among laboratories.

The Panel summarized the general grading scores/severity scoring system in Appendix 1. During the meeting, the Panel questioned why morphometric analysis was not used instead of severity scores when the percentage of affected tissue is known. The grading/scoring system is a non-quantitative method for histopathology analysis and is not easily conducive to statistical analysis as the data are not continuous.

A complicating factor in all of the histopathological analyses was insufficient replication, particularly in view of the loss of some samples. For instance, the study report entitled, *Histopathology Evaluation of Reproductive and Selected Visceral and Brain Tissue from the Avian 2-Generation Test in Vinclozolin Exposed Adult Japanese Quail*, stated that “the pineal gland was present in only 2 of 10 F1 males and 1 F1 female.” The Panel noted that all birds have pineal glands and these may be attached to the cranium and thus unless care is taken, they can be lost during dissection.

The Panel noted the lack of consistent effects of vinclozolin on histology for most tissues examined (see Appendix 2). For instance, the study report (Troan *et al.* 2012) stated that samples from Laboratory 1 had “no lesions attributed to vinclozolin exposure in the adrenal gland, heart, liver, kidney, brain (pre-optic nucleus-POM, pituitary or pineal gland), and thyroid gland of male or female birds”. In contrast, Laboratory 3 results indicated there was a consistent effect of vinclozolin on the liver, “...minimal to mild, random multifocal vacuolization was observed. Hepatic vacuolization was characterized by random and multifocal hepatocytes.” Adverse effects of vinclozolin on the epididymis were consistently observed in the inter-laboratory studies. The Panel indicated that these results were consistent with known anti-androgenic effects of this putative endocrine disrupting chemical. All three laboratories showed an increase in the incidence and severity of efferent duct dilation in F0 and F1 adult males exposed to vinclozolin.

Recommendations

The Panel made the following recommendations to improve the histopathology protocol and its execution.

- 1) Gonadal, epididymal, liver, kidney, and thyroid histopathology should be part of the JQTT protocol.
 - a. Histopathology of the gonads allows determination of full or partial sex reversal.
 - b. The epididymal histological data, although limited, are consistent within each and between laboratory and produce results that are consistent with the expected anti-androgenic effects of vinclozolin.
 - c. Thyroid histopathology provides evidence for thyroid toxicity (Dean *et al.*, 1991; Gentles *et al.*, 2005; Sonne *et al.*, 2010; liver: Rattner *et al.*, 2011; 2012; Magnoli *et al.*, 2012; kidney: Jayakumar *et al.*, 2010; also see review: McNabb, 2007). Circulating concentrations of the

active thyroid hormone, triiodothyronine (T3), are controlled by availability of thyroxine from the thyroid and liver moniodinases, either activating or inactivating T3. Another very useful biochemical endpoint for the thyroid is liver monodeiodinases. Polychlorinated biphenyls have been demonstrated to influence type 1 monodeiodinase in the chick embryo (Gould *et al.*, 1999)

- d. Liver and kidney histopathology provide considerable evidence that toxicants exert profound effects on these organs (liver: Rattner *et al.* 2011, 2012; Magnoli *et al.* 2012; kidney: Jayakumar *et al.* 2010; also see review: McNabb, 2007).
- 2) All treatments should include a positive and negative control with “blind” treatments. The choice of the chemical to be used for a positive control should be made based on the effects of the putative endocrine disrupting chemical in the molecular/cellular screens. For instance for a putative estrogenic endocrine disrupting chemical, estradiol would be a suitable positive control or for a putative androgenic endocrine disrupting chemical, e.g., testosterone or trenbolone, would be a suitable positive control. The inclusion of a positive control will demonstrate the validity of the results particularly when there are no effects of the test agent. Moreover, the presence of both a negative and positive control will provide a quality control for the contract laboratories. The treatments are to be tested “blind” to assure removal of potential bias and subjectivity. This is consistent with the general recommendations stated at the beginning of the report.
 - 3) EPA should use a quantitative analysis as a means to enhance analytical power and consistency among laboratories. Where possible, histopathology studies should be conducted using a blind approach coupled with morphometric analysis. A quantitative histopathology analysis would use a systematic series of sections for each tissue conducted using analysis of sections either in a systematic or random basis. In the systematic approach, every 3rd or 5th (or some specific number) section is reviewed. In the random approach, randomization is used to select which sections are reviewed. Image analysis can be readily automated. The Panel indicated that these improvements will reduce bias and facilitate statistical analysis since the data will be continuous and therefore amenable to the use of parametric statistical analysis. This, in turn, will increase consistency in performance and analysis of the histopathology.
 - 4) At least eight random samples (per treatment) for each organ should be subjected to histopathological analysis to provide sufficient statistical power. The present protocol requires “five replicates per treatment (for adult males or females or embryos) for histological analysis” but even with this low number of specified replicates, the actual number of tissues sampled in the inter-laboratory studies was often lower.
 - 5) EPA should research the applicability of immunocytochemistry/*in situ* hybridization approaches as additional endpoints. These methods are much more state-of-the-art and provide evidence of mechanism of action.

Question 4. Demonstration of the test method performance should be based on the testing of reference chemicals representative of the types of substances for which the test method will be used. Test substances should adequately represent an appropriate range of responses and physical/chemical properties for which the test method is proposed to be appropriate. The selection of the most appropriate statistical approaches depends in part on the nature of the data and also on the design of the validation study. Statistical and non-statistical methods used to analyze should be described. Please comment on the selection of test substances and methods (analytical and statistical where appropriate) chosen for the demonstration and validation of the JQTT assay.

Panel Response

- 1) Test substances.** The Panel indicated that trenbolone was a very informative choice for a test substance, but raised several concerns about the choice of vinclozolin as a central component of the inter-laboratory tests. In particular, the lack of a strong *a priori* expectation of how vinclozolin should impact birds made it difficult to interpret the negative results obtained from the inter-laboratory study. This again emphasizes the need for appropriate positive controls and a formalized system of reference testing for labs to demonstrate proficiency with the protocol.

The Panel emphasized that endocrine disrupting substances which introduce biological effects on the mammalian endocrine system might affect the avian endocrine system differently, or not at all. There are often quite large differences between mammals and birds in terms of protein structure of hormone receptors. Thus, a chemical that cannot bind to a mammalian hormone receptor and has no endocrine disrupting effects in mammals might well bind the equivalent (or different) hormone receptor in birds, and vice-versa. Thus, the current methods employed in Tier 1 testing might not be appropriate for non-mammalian species, and might provide false positive results of chemical screening or false negatives. This might be a reason that the test chemical employed in the inter-laboratory tests, vinclozolin, produced no treatment effects. Use of a different chemical, perhaps trenbolone, as a positive control, would have helped interpretation of the data from the inter-laboratory test (see comments below under heading, “Data interpretation and need for positive control”). Further, birds that raise altricial young, such as the passiformes, might respond differently from Japanese quail to the test substance. Thus, the Panel strongly recommended (as elsewhere in this document) that zebra finches, songbirds that raise altricial young, also be included in this test.

- 2) Analytical Methods.** The Panel recommended that a combination of standard laboratory molecular biology tools, e.g., Real-Time Quantitative PCR (RT-qPCR or qPCR), and immunohistochemistry (IHC) techniques should be used to analyze the potential effects of any test chemical on neuroendocrine and endocrine tissues. The combination of these techniques provides an extremely powerful analysis of the tissues to be investigated.
 - RT-qPCR or qPCR should be used for some of the endpoints in addition to histology. The Panel noted that a very clear and specific SOP should be developed for inclusion of this method in the protocol and used across all laboratories. Some points that should be addressed

in the SOP include the specific qPCR conditions, primers, reference genes and software for data analysis.

- IHC techniques, although challenging, should be used in conjunction with qPCR and the stains already employed. IHC will require validated antibodies and morphometric analysis.

In combination, these techniques will provide information on the endocrine disrupting chemical effects of any test chemical at the molecular, hormonal and structural levels. This thorough analysis, combined with a behavioral analysis, will increase the robustness of the assessment for any test chemical. Additional discussion on analytical techniques is found in the Panel's response to charge question 6.

3) Statistical Methods. The JQTT ISR and background documents clearly describe what can be controlled, but not enough concerning how randomization was used. The Panel recommended a detailed discussion of randomization be included in the revised protocol. The following questions illustrate the information needed to describe how randomization was used in the experimental design. The answers to these questions become part of the detailed description of the experimental design in the protocol.

- Are all caged pairs located in the same room or in different rooms?
- If they are in the same room, where are they located in the room?
- Are these locations selected randomly?
- If they are located in different rooms, are the rooms assigned to the caged pairs randomized?
- Are the cages identical to one another?
- Are the feed and water containers placed in the same location in each of the cages?
- Since one cannot place all feed and water containers simultaneously, is the order of placement to be randomized? If not what order is to be used?
- Are all of the birds in a given treatment group to be housed together? If so, this would invalidate all of the statistical analyses that are being performed as treatment effects would be confounded with housing effects.
- Are all dose levels done at the same time? If not, how is their order to be randomized?
- How is randomization to be used to select pairs to be mated?
- Are eggs that are chosen to be evaluated selected randomly?
- Were observation times during the day randomized?
- How are the initial birds chosen for each of the laboratories?

The Panel also provided comments and recommendations for specific items within the protocol that should be addressed in the JQTT protocol.

- The protocol states that if “*the dose level during the 6th and 7th post-treatment weeks does not produce at least 40 eggs, then that level is dropped from further analysis in the test. It is dropped from further testing.*” Are the 40 eggs for all pairs together or for a single pair?
- The protocol states that “*at 4 weeks of age, F1 control and treatment group birds are weighed and randomly allocated to pens by pairs with male offspring from odd numbered parental pens being paired with females from even numbered parental pens and even numbered pen males being paired with odd numbered females.*” Does this guarantee that paired birds do not come from the same lineage? Does this matter? The Panel suggested that the birds be paired by a restricted, stratified randomized process so that birds from the same parents are not paired as stated in the protocol: “*The process should be a stratified random process to assure, to the extent possible, that each of the F0 pairs are contributing genetically to the F1 breeding population in order to represent individuals that may vary in sensitivity to chemical exposure.*”
- The protocol states that “*The frequency of diet renewal should not be more than once a day and not less than once a week.*” The Panel recommended that the diets be renewed daily to prevent degradation of the test agents. Also, food and water should be withheld from the birds for about 4 hours prior to any weight measurements and blood or tissue samples being taken. Moreover blood samples should be taken at a consistent time of day.
- The protocol states that “*Incubators and hatchers, preferably with automatic temperature and humidity controls and an egg-turning device, are necessary. In addition, suitable equipment is required to maintain stored eggs within the temperature and humidity ranges specified.*” Is the incubator for each pair of birds or all eggs from all birds placed in a single incubator? If a single incubator is being used, are the locations of the eggs within the incubator randomized? Are the eggs from each pair kept together in the incubator(s) or randomized throughout the incubator? The Panel recommended there should be randomization and this should be explicit in the protocol.
- The protocol states that “*Chicks should be identified individually or by pen of origin. They may be housed in groups of approximately equal number, by treatment group.*” The Panel indicated that it is acceptable to house chicks from the same parent pair together, but was concerned with possible mixing of chicks from different parents together even though their parents received the same treatment dose. The Panel pointed out that “convenience” of housing chicks from multiple parents together is not good experimental design.
- The protocol states that “*All F0 birds used in a test should be from the same hatch.*” The Panel noted that F0 eggs should come from the same hatch, but with multiple pairs.

However, as three generations of quail are being used, the protocol should clearly state that the F1 or F2 generation birds are not the result of brother-sister mating.

- The protocol states that *“The eggs are stored in a cold storage facility, for a maximum of 2 weeks, prior to setting in the incubator”* (see Table 3-3). The protocol did not clearly explain whether the eggs were placed at random locations in the cold storage facility and whether the eggs from the same pair of parents were kept together. These points should be stated explicitly in the protocol.
- The protocol states that *“During the study, one egg per pen is collected from odd numbered pens in odd numbered weeks and from even numbered pens in even numbered weeks to evaluate eggshell strength and thickness.”* Is the egg randomly selected from those eggs available? If so, how is randomization assured?
- The protocol states that *“Fertility and embryo viability in all incubated eggs are determined by candling the eggs after 8 days of incubation (ED 8) and 2 days before hatching at embryonic day 15 (ED 15), eggs will be transferred to a hatcher.”* The Panel stated that randomization should be involved in this process. It is important to be able to identify which chicks come from which eggs to ensure randomization.
- The protocol states that, *“Some sample sizes were too small for statistical analyses.”* The Panel recommended that sample sizes be increased to allow for meaningful statistical analyses.
- The protocol states that, *“Often, there have been non-linear responses observed in these endpoints; however, this is sometime due to the number of individuals examined and could be assessed through a power analysis to determine the vigor of that particular endpoint.”* See the Panel’s overall recommendations for the experimental design and statistical analysis.
- The protocol states that, *“Comparison of the data from the three laboratories shows evidence for the following: There are clear differences in strains of Japanese quail in terms of body weight, and related variables that contribute to somewhat large differences and ranges in measurement endpoints. Strain differences may also be a factor in rate of maturation and reproductive characteristics.”* The Panel commented that different strains used by different laboratories might be a compounding effect. In response, the Panel recommended that the protocol should clearly state the strain of Japanese quail used in the assays.
- The protocol states that, *“Power determinations were carried out for each endpoint based on the control group within run variance.”* The Panel questioned the meaning of “within run variance” used in this circumstance.

*Question 5. Considering the variability inherent in biological and chemical test methods, a test method needs to be repeatable and reproducible. A test is robust and reliable if the results are repeatable and reproducible within a laboratory and between different laboratories, respectively. A test protocol should provide sufficient guidance to ensure proper and consistent performance across labs and chemicals. **Please comment on the test method robustness and reliability and the repeatability and reproducibility of the results obtained with the JQTT assay.***

Panel Response

The Panel concluded that several of the tests have high CVs and low power to detect changes; however, there was not enough detail to distinguish between variability introduced by the test protocol and inherent biological variability. If the variability is a function of laboratories using slightly different procedures, then this can be corrected by being more specific in the protocol technical guidance and strict adherence to it. However, if the variability reported reflects biological variation, then the sample size may need to be increased.

The Panel provided the following specific comments and recommendations to improve the test method robustness and reliability and the repeatability and reproducibility of the results obtained with the JQTT assay.

- 1) Data interpretation and requirement of positive control.** The Panel strongly recommended the use of a positive control chemical with known endocrine disruptor activity should be included in all experiments wherever possible. This would allow for better analysis of the reproducibility and repeatability of the test. For example, the inter-laboratory study used a single reference chemical, vinclozolin. None of the laboratories showed much in the way of treatment effects in this test. As it stands, the Panel could not determine the reliability and repeatability of the results obtained. The use of a strong AR agonist trenbolone, for example, would be a suitable positive control substance. A single dose at a concentration known to induce a clear adverse outcome in some, or all the endpoints included in the JQTT, would suffice. Without the use of a positive control, it is impossible to evaluate datasets with a negative outcome, especially given the variability in data from different laboratories. Although it may seem that including an extra group in the study will add to the overall cost, there will be cost-saving in the long run as use of a positive control will provide solid information as to the reliability of any future negative results, negating the need to repeat experiments.
- 2) Strain differences.** The Panel observed that a single strain of Japanese quail was not used in these studies. The ISR was unclear in identifying the strains used by each laboratory, but the Panel noted two of the three laboratories obtained their Japanese quail from a single breeder (Northwest Gamebirds, Kennewick, WA) and the third laboratory bred their own Japanese quail. The Panel indicated that using different strains of Japanese quail in the inter-laboratory vinclozolin study is a potential source of variation in the data collected. The ISR also noted the same point (see lines 29-44, p. 64), “*there can be marked inter-strain differences in body mass, endocrine characteristics, fertility and sexual behavior.*” The OECD JQTT guideline document

(OECD, 2002) states that, “*Many strains of Japanese quail have been developed, largely along egg production or body mass lines. The impact of strain selection on the ability of the test to detect endocrine activity and reproductive deficit needs to be evaluated.*” The Panel echoed the OECD’s recommendation. The Panel also commented that genetic drift caused by continued colonization in the laboratory might also be a factor causing variation in the inter-laboratory data.

The Panel recommended that a standard Japanese quail strain be used by all laboratories. This is vital to the repeatability of the assay. The Panel suggested founding a single genetic stock that is housed in more than one location (in case of disease or natural disaster) and have appropriate genetic husbandry performed between these locations periodically to minimize genetic drift. This stock can then be sent to the test laboratories as needed. The Panel suggested that the stock strain be made available internationally.

- 3) **Egg storage and incubation parameters.** Egg storage and incubation metrics (humidity, slight temperature effects on hatchability) do not appear to be completely standardized or factored into the analysis. The Panel stated that some variation in the inter-laboratory results likely arises from subtle differences in incubation parameters employed between laboratories. Humidity, incubation temperature and egg storage temperature can all have large effects on hatchability. Housing light intensity should be stated in the protocol and standardized across and within labs. All birds should be exposed to the same light intensity under a given photoperiod.
- 4) **Time of sampling.** The Panel recommended that a standard time of day of sampling within and between laboratories (mass, hormone measurements) be established.
- 5) **Light intensity.** Light intensity differences in housing do not appear to have been accounted for. When exposed to a particular long day length (e.g., 18h light per day), photoperiodic birds will exhibit different reproductive responses according to light intensity. Under low light intensities, photoperiodic birds will exhibit reduced photoperiodic responses compared to birds exposed to the same 18h day length but under high light intensity. There were different light intensities employed in the inter-laboratory studies: 77 lux (Wildlife International study), 10 lux (Smithers Viscient study), and the Bayer study simply states that, “*a low light intensity was utilized to calm the birds and reduce stress.*” Even where light intensity at the front of the cage was reported, the use of racks with tiers of cages (as seems to have occurred in the Bayer study) compounds this issue, as birds housed in lower tiers of cages will experience lower light intensity than birds on the top rack of cages. The Panel recommended that a standard housing light intensity should be clearly specified in the protocol for use by all laboratories. All birds should be exposed to the same light intensity in a given photoperiod.
- 6) **Stress lines within a strain and need to measure circulating corticosterone.** Hypothalamic-pituitary-adrenal responsiveness within a strain was not measured to ensure a normal distribution of stress response. In some quail strains there are “high stress” and “low stress” lines with highly different physiological and behavioral responses to a stressor (e.g., Satterlee and Roberts, 1990; Satterlee and Marin, 2006; Satterlee *et al.*, 2007; Cockrem *et al.*, 2012; Huff *et al.*, 2013). This

differential response could be a source of variance in the data. Low stress lines vs. high stress lines may well respond differently to endocrine-disrupting chemicals that interact with corticosterone binding globulin (CBP) and cause an effect on circulating gonadal steroids. This difference in stress-induced corticosterone release would most easily be detected via blood sampling using a standardized stressor protocol given below. Also, baseline circulating concentrations of corticosterone will likely differ as a result of differences in secretion or clearance in response to treatment. Sampling should occur at the same time of day for all birds. The Panel recommended measurement of the corticosterone response to standardized capture-handling stressor to ensure equivalent response across groups and no bias as result of lines with different stress responsiveness. Here, corticosterone would be more easily measured in serial blood samples and subsequent radioimmunoassay or ELISA, rather than from fecal samples.

- 7) **Histology: tissue selection.** The Panel stated that clear guidance on the selection of animals, tissue, and sections for histological and subsequent statistical analysis should be developed. At present this is vague. The Panel stated that there needs to be clear guidance on the selection of animals, tissue, and sections for histological and subsequent statistical analysis. Currently, the protocol provides the following description for selection of tissues for histopathological analysis (p.6, EPA-HQ-OPP-2013-0182-0069): *“Images of representative exposure-related effects and associated severity scores were prepared to be used as references for scoring the remainder of the specimens. After the initial evaluation, the study pathologist performed a masked re-evaluation of all specimens. The single most representative section was histologically evaluated from each tissue.”* This description is inadequate. For example, what is a “representative” section? What are the selection criteria for which individuals were selected for tissue analysis and when the tissue was excised, what criteria were used to select specific section for histological analysis? Was selection random for animals selected from each group for histological analysis? If sample size is low, what is the rationale for not using all of the animals? The Panel recommended that histological assays (tissue selection, quantitative method, preferably morphometric analysis) be standardized. Tissue sections should be selected on a randomized basis or preferably, a systematic basis, e.g., every 3rd, 5th and 7th section throughout the tissue.
- 8) **Behavioral assays.** The Panel commented that standard operating procedures for behavioral assays are not always as detailed as for biochemical assays. Changes in behavior are often rapid, subtle, and difficult to measure and require more detailed explanation and training to perform the assay as expected. Measurement of behavioral endpoints can be highly variable because of differences in the level of experience conducting the assay. Behavior is an important endpoint in birds because previous research established a clear link between endocrine disruption as a mechanism and effects on behavior. Behavioral changes can be linked to those “fitness” endpoints incorporated into an ecological risk assessment. The Panel recommended a more detailed and standardized SOP be developed to increase transparency in how the assay should be conducted and technical guidance/training to increase the proficiency in performing the assay. Behavioral assays are highly variable and greater standardization and technical guidance would assist in reducing inter-laboratory variability.

- 9) **Sample size.** The Panel recommended that sample size be increased for measurements of all endpoints, particularly hormone sampling, tissue samples and behavioral assays. This would be especially important if adherence to stricter and standardized operating procedures does not reduce observed variability.

Question 6. The test protocol should be descriptive enough to be fully transferable to a competent laboratory. The protocol should describe the methodology of the assay in a clear and concise manner so that a laboratory could comprehend the objective, conduct the assay, observe and measure prescribed endpoints, compile and prepare data for statistical analyses, and report the results. Section 6 of the JQTT ISR outlines the process of and challenges experienced in the inter-laboratory studies. Please comment on the transferability across labs and provide any suggestions or recommendations for improvement of the JQTT assay.

Panel Response

The Panel stated that the test protocol is not descriptive enough to be fully transferable across laboratories. Data from the inter-laboratory studies are highly variable across many endpoints. Variability among laboratories may be due to poor conduct of the assays, lack of clear and concise descriptions of assay methodology, lack of standard operating procedures, lack of guidance/training for conducting the assays, and choice of statistical methods/analysis. See also response to charge questions 5 and 7. The Panel recommended that a very clear and detailed set of standard operating procedures including instructions on intra- and inter- laboratory standards be more fully developed and included in a revised protocol.

Recommendations

The Panel had the following recommendations to improve the transferability of the JQTT:

- 1) Provide more detailed and clearer standard operating procedures for all assays (biochemical, histopathological, and behavioral);
- 2) Include positive controls representing different endocrine disruptor pathways for each assay to be run alongside each test chemical to judge consistency in the assay performance, increase certainty in understanding negative results, and allow objective analysis of data from different laboratories;
- 3) Distribute the same biochemical standards to all laboratories as internal controls for the assay;
- 4) Standardize the Japanese quail strain, i.e., use a single source that has appropriate husbandry to prevent genetic drift and ensure that a random-bred population is used. See also discussion in the Panel's response to charge question 8. Standardize egg storage and incubation metrics (humidity, temperature);
- 5) Standardize the time of day blood and tissue are sampled;

6) The Panel made a number of specific recommendations concerning the behavior tests including the following:

- Standardize the time of day blood and tissue sampling occurs and when behavioral tests are performed.
- Expand and standardize the behavioral assays to include appetite and consumption behaviors;
- If a passerine species is to be used as part of this test (see response to charge question 1), then courtship behaviors (directed singing versus undirected singing) needs to be measured. Importantly, female sexual behavior needs to be included as part of this test. For quail, use of a female choice chamber might be suitable, but female sexual behavior would be more easily measured in zebra finches. Parental care needs to be measured, as it can have population-level effects. For quail, egg incubation and attention to hatchlings can be measured. Again, parental care would be more easily measured in a passerine species (e.g., egg incubation, rates of feeding of hatchlings).
- Provide training for conducting and interpreting the behavioral tests.

7) Histology

- Provide a clear rationale for selection of individuals for histopathology;
- Provide a detailed standard operating procedure for scoring and statistical analysis; and,
- Increase the number of replicates to at least n=8.

8) Demonstrate proficiency in conducting each assay. The Panel strongly recommended the use of guidance manuals and training sessions, e.g., hands-on workshops, virtual training, and preliminary testing with representative chemicals, to improve the reproducibility and reliability in the data. If laboratory personnel increase their proficiency in performing the assays according to the specific instructions in the protocol than this should result in more consistent data and provide cost-savings in the long run. Training programs would also provide a mechanism to provide feedback on the clarity of the protocol, effectiveness of the training program and test the laboratory's performance. Laboratories running these protocols over time might identify portions of the protocol that might need to be clarified (written instructions) or revised (number of positive controls, endpoints, etc.) without having to use a large number of animals.

*Question 7. The purpose of the validation process is to determine the readiness of a test for inclusion in a testing program. A component of readiness of a test is the evaluation of the usefulness and limitations of the test, including the classes and types of test substances that can and cannot be tested. **Please comment on the strengths and/or limitations of the JQTT assay, as described in Section 7 of the JQTT ISR.***

Panel Response

The 1998 EDSTAC recommended a two-tiered endocrine disruptor testing system. Included in the EDSTAC recommendations were the following Tier 2 tests: mammalian reproductive toxicity test, an avian reproduction test, fish life cycle, amphibian development and reproduction life cycle tests and a mysid (invertebrate) life cycle test. The proposed Tier 2 avian reproduction test was developed using the OECD guidelines for the Japanese Quail Two-Generation Reproduction test (see Appendix D, JQTT ISR). The Panel noted that the OECD originally considered two species; Japanese quail and Bobwhite quail, and opted for Japanese quail as the model species:

“Although the Japanese quail is not indigenous to the United States and has undergone extensive domestication, and therefore may be less representative of wild species, it is recommended here as the preferred test species because of its small size, high fecundity, well-characterized reproductive biology, and in particular, it’s very rapid incubation and maturation stages.” (See Appendix D, JQTT ISR.)

Strengths

- 1) The logic supporting the use of Japanese quail as the chosen species is reasonable. The Japanese quail offers a series of specific advantages summarized in the JQTT ISR. In general, Japanese quail is a solid pragmatic choice as a surrogate bird. Japanese quail have a short generation time and are readily reared in the laboratory. The species has been domesticated and has a long history in captivity. For this reason, Japanese quail are much easier to work with in the laboratory than the alternative species, Bobwhite quail (OECD, 1996; 2005; 2007; 2009; 2011). These practical considerations are extremely important, especially given that the JQTT will need to be adopted by a variety of laboratories with differing degrees of experience with it.
- 2) There is also considerable published research on endocrine physiology of Japanese quail, particularly reproductive and thyroid endocrinology, which will help place results in the context of other laboratory studies.
- 3) The JQTT measures key endpoints: egg production, growth, embryo development and phenotypic sex (versus genotypic sex). Decreases in egg production or growth or impaired embryonic development or sex reversal meet the criteria of reliability and relevance. The effect of putative endocrine disruptors on egg shell thickness is also a critical endpoint.
- 4) The Panel agreed with the following OECD conclusion, i.e., the Japanese quail is moderately sensitive to putative endocrine disrupters (OECD, 2007).

Limitations

1) Japanese quail as a representative species has limitations

The Panel spent considerable time in discussing the limitations of Japanese quail in response to charge question 1. The practical factors that make Japanese quail an attractive

species for this test, also make it more challenging to extrapolate from Japanese quail to wild bird populations that may be the subject of environmental risk assessments. Most birds are altricial, most are in the (Super Order: Neoaves, not in the evolutionary side-branch with Japanese quail (in Galliformes). In their recent book chapter, Jones *et al.* (2013) point out limitations of depending on Japanese quail in this test and emphasize that Japanese quail are not representative of many birds.

While many characteristics of the endocrine system are conserved across related species, significant differences do exist among bird species in both endocrine pathways and in the life history traits that will determine the effects of endocrine disrupting chemicals on populations (Adkins-Regan 2008, Adkins-Regan *et al.* 2013, Jones *et al.* 2013, Ottinger *et al.* 2008, Starck & Ricklefs 1998). As an example, the large number of eggs laid by Japanese quail females and the fact that precocial young can be raised with little intervention from parents or laboratory managers are practical advantages. However, the majority of bird species, and all passerine species (*Passeriformes*, the largest order of birds), are altricial, hatching earlier in development and dependent on extensive parental care to survive and develop. While there is no evidence to support the idea that either precocial or altricial species will be more sensitive to the effects of endocrine disrupting chemicals, the Panel recognized that there may be differences in how precocial and altricial species respond to endocrine disrupting chemicals. An obvious example of the differences would be the importance of parental care in altricial species. While parental care does occur in precocial species, these behaviors are not evaluated in the JQTT and are by definition more pronounced in altricial species. Parental care behaviors include incubation, brooding, feeding and other behaviors and are known to be regulated by the endocrine system and are thus vulnerable to endocrine disruption (Clotfelter *et al.* 2004, Zala & Penn 2004). Parental care behavior is critical to successful reproduction and integral to population dynamics and therefore may be a key endpoint in evaluating population level responses to endocrine disrupting chemicals.

The Panel stated that the JQTT appears to be a highly appropriate to test potential endocrine disrupting chemicals in Galliform species with precocial chicks, but indicated that there is uncertainty as to how applicable it will be to wild avian species. Long-lived species such as albatross and gulls take several years to enter into puberty. As a result, they potentially experience a much longer exposure to endocrine disrupting chemicals than Japanese quail. This delayed entry into puberty might make them more or less sensitive to the chemicals tested. A passerine species with altricial young might be more or less sensitive to endocrine disrupting chemical than Japanese quail. The altricial young hatch in a highly undeveloped state (blind, naked) and are often fed insects by their parents. These could be considered “external embryos”. Thus, bioaccumulation of endocrine disrupting chemicals in their food combined with exposure during critical developmental stages might make them more susceptible to endocrine disrupting chemicals than birds with precocial young, such as Japanese quail. Overall “fitness” may be more easily measured in songbird model with more pronounced and hence readily observed and quantified secondary sexual

characteristics. Thus, sexual selection could be measured more easily than in quail. In addition, parental care could also be measured more easily.

2) Intra- and inter-laboratory variation

There were high coefficients of variance (CV) in the control birds for a number of important endpoints for controls between and within the different laboratories that will limit the power to detect differences. Lack of clarity in the protocol and lack of proficiency in conducting the JQTT in the laboratories due to lack of appropriate training, are key sources of variation. The Panel elaborated on other possible sources of intra- and inter-laboratory variation.

- There were large differences in Japanese quail body weight between the different laboratories. This may be attributable to strain or other genetic differences.
- There was a lack of consistency in egg shell thickness determinations between the different laboratories and within a laboratory between generations. The availability of appropriate intra- and inter-laboratory standards, such as the use of a positive control, would reduce inter- and intra-laboratory variability.
- There is inadequate replication in the current protocol. The Panel recommended increased replication and referred to the OECD guidelines (2007; 2011) for support, *“in the OECD draft guideline, the initial test groups would consist of 20 replicate pens to increase the likelihood that at least 16 of them remain in each group at test termination.”* Moreover, OECD recommended determination of *“Sperm Motility and Morphology, and Fertilization Success”* circulating corticosterone (the major glucocorticoid in birds) and VTG. These inclusions in the protocol are either not done or done very sparingly.
- Lack of morphometric analyses in the histopathology studies and concerns with the statistical approaches used are addressed in the general overview at the beginning of the report and in charge questions 3.2 and 4, respectively.
- The use of F0, F1, and F2 terminology was confusing and needs to be clarified.

Recommendations

- 1) **Defined population of Japanese quail.** There is a critical need for a defined population of Japanese quail available on a widespread (optimally global) basis to test laboratories. This is consistent with the recommendations of OECD (2002), *“Many strains of Japanese quail have been developed, largely along egg production or body mass lines. The impact of strain selection on the ability of the test to detect endocrine activity and reproductive deficit needs to be evaluated”*. The Panel encouraged the Agency to lead efforts to establish, standardize and oversee such a population.

- 3) **A zebra finch two generation reproduction test.** The Panel discussed this recommendation in great detail in response to charge question 1. The Panel expressed their concern for extrapolating the results of two generation reproduction test using Japanese quail to assessment of the risk to wild birds. The difficulty in extrapolating to other avian species is not limited to long-lived species as noted in the ISR. The Panel recommended the use of a two generation study with a passerine species, e.g., zebra finch.
- 4) **Detail of protocol.** The protocol should provide sufficient detail to ensure that the experimentation is performed correctly and consistently between laboratories.
- 5) **Egg shell thickness and behavioral endpoints.** The effects of putative endocrine disrupting chemicals on egg shell thickness and behavioral endpoints are unique endpoints that should be maintained as part of the JQTT.
- 6) **Epigenetic endpoints.** The Panel recommended that the Agency pursue research on epigenetic endpoints. See response to charge question 3.1.
- 7) **Replication and statistical approaches.** Replication should be increased. Concerns about the lack of morphometric analyses and about the statistical approaches are addressed in the response to questions 3-2 and 4, respectively.
- 8) **Training programs.** Training programs are critical to understanding the protocol and performance of the JQTT. Laboratories performing the JQTT should have sufficient training such that the laboratories can demonstrate proficiency in the JQTT. A standardized set of training programs (e.g., webinars, web applications, workshops) should be developed and implemented across all laboratories to increase proficiency in running the tests and increase the intra- and inter-laboratory reliability and reproducibility.
- 9) **Clarify terminology.** The use of F0, F1, and F2 terminology is confusing and needs to be clarified. Future documents should clearly state that the F0 generation is the generation in which treatments begin.

Question 8. There is sufficient evidence to indicate that endocrine disrupting chemicals can disrupt normal development and reproductive success, however the sensitivity of the F2 generation compared to the P0 or F1 is less clearly defined. In the JQTT, there is an increase in endpoint CVs with each subsequent generation (JQTT ISR Table 6-4), which indicates decreased power of discrimination. Please comment from a scientific and risk assessment perspective on the value added of multiple generations in the JQTT assay.

Panel Response

The Panel concluded that multiple generations in the JQTT added significant value to the test results from both scientific and risk assessment perspectives. The F2 generation, where exposure occurs solely during embryonic development, either provides or is capable of providing critical information as to the endocrine disruptor effects of pesticides and other chemicals employed in the environment. Increasing evidence for cross-generation effects emerging from research labs as well as trenbolone results

presented in the JQTT ISR suggest that the current JQTT may undervalue the importance of endpoints measured in later generations. Overall, the Panel expressed concern that the statistical power of the current JQTT may be too low. Small sample sizes make it difficult to detect biologically and ecologically important differences, if they exist. The Panel recommended that the number of replicates should be increased in addition to testing multiple generations.

Results emerging from research laboratories increasingly suggest cross-generation effects may be important consequences of endocrine disruption and that these types of effects are expected to occur (Frésard *et al.* 2013, Groothuis *et al.* 2008, Lewis *et al.* 2012, Hala *et al.* 2012, Meylan *et al.* 2012, Zhang and Ho 2012). As research continues to point in this direction, negative results from evaluation of F2 birds become an important contribution to risk evaluation. Without data on the F2 generation, the test information will be incomplete and risk assessment results would be assumed to underestimate risk to bird populations by not incorporating cross-generation (and epigenetic) effects.

The following comments refer to slides in the EPA presentation made during the meeting (a copy of the presentation is found on the public e-docket website: <http://regulations.gov>, docket number, EPA-HQ-OPP-2013-0182-0078). The trenbolone tests of the number of eggs laid per day within each set interval for the F0, F1, and F2 generations (slide 37, U.S. Geological Survey study) indicated that trenbolone had a significant effect on egg production in the F2 generation. This result is in contrast to the finding for vinclozolin, which showed no detectable effects in the F2 generation. The U.S. Army testing laboratory study (Appendix E.3, JQTT ISR) results for average egg production per week for the F0, F1, and F2 generations (slide 38) also indicated F2 generational effects. In the U.S. Army test (Appendix E.3, JQTT ISR), results suggest that the effects of trenbolone exposure of the F0 generation increases the sensitivity of F1 females to endocrine disrupting effects of this chemical. The Panel stated that these finding underscore the importance of performing dose studies with F2 animals to determine if lower doses might impact the reproductive function with longer intergenerational exposure.

The proposed protocol does not include an endpoint for F2 reproduction. The Panel indicated that if reproductive effects in the F2 generation is a standard response to exposure to an endocrine disrupting chemical, or if exposure increases sensitivity to endocrine disrupting chemicals across generations (as suggested by the trenbolone results), then inclusion of F2 endpoints will be vital to informing wildlife population models used in an ecological risk assessment. Modeling will also identify vulnerability of various life stages to exposure.

The Panel commented that the charge question implies that the answer to the question about the value of the F2 in the JQTT is to be found by simply looking at the patterns of CVs across generations. As stated in the charge question, *“the sensitivity of the F2 generation compared to the P0 or F1 is less clearly defined. In the JQTT, there is an increase in endpoint CVs with each subsequent generation (JQTT ISR Table 6-4) which indicates decreased power of discrimination.”* The Panel disagreed with the EPA’s assumption that an increase in endpoint CVs across generations is indicative of lower value. If the pattern of increased variation described by the CV is actually biological and not an artifact of the way the tests were conducted, then this is a biological effect of interest. If within a test, variability does increase in offspring relative to their parents, this would confirm the need for the multi-generation test. The phrasing of the charge question confuses the necessity for the test with the statistical power needed

to meet the experimental design objectives. If exposure to endocrine disrupting chemicals does result in greater variability among the offspring of exposed individuals, then the limited statistical power associated with the greater CV becomes a sample size and design issue. The protocol would need to be revised to add more breeding pairs to ensure relevant differences can be detected.

The Panel did not accept the flawed premise that an increase in CV across generations would indicate that the F2 step is not a valuable addition to the protocol. Moreover, the results of the inter-laboratory studies do not support the conclusion that endpoint CVs are increasing across generations. The pattern of variation in CVs does not seem straight-forward, the methods of determining CV are unclear; therefore, the Panel was unable to draw definitive conclusions about the sensitivity of the F2 generation from those results. It is clear from examining Table 6.4 in the JQTT ISR that the CVs are test-specific. The “total” CV of F2 is indeed higher, but some of this is because the testing of the F2s included two variables with high variation all generations (biochemistry and behavior) and did not include several variables with low variation in the other generations (reproduction). There may be valuable insights to be obtained from analyses of the differences in these variables, but the summary report was insufficient for the Panel to come to a conclusion.

One argument presented against inclusion of the F2 generation was the suggestion that the inter-laboratory tests with vinclozolin showed little or no evidence of effects on F2 birds. That said, members of the SAP expressed concern that the inter-laboratory test results on vinclozolin presented may not characterize the potential for adverse effects to be exhibited in the F2 generation. Critiques presented elsewhere in this report raise the concern that the lack of differences seen in the F2 generations is driven by low statistical power or other problems with the protocol or the application of the protocol. These concerns emphasize the need to strengthen the protocol so that risk assessors have the ability to distinguish “true” negative results from “false” negatives that would prove to be misleading for risk assessment.

Overall, the Panel expressed concern that the statistical power of the current JQTT may be too low. Small sample sizes make it difficult to detect biologically and ecologically important differences, if they exist. The Panel recommended that the number of replicates should be increased in addition to testing multiple generations. Some panelists discussed a scenario that would increase statistical power by increasing sample size in the F0 generations while eliminating F2 endpoints. Under this scenario, the dose strategies and the choice to continue treatment through adulthood in the F1 would need to be modified. This scenario illustrates the difficult balancing act between the value-added of multiple generations versus increased sample sizes to increase power of the test. The Panel recommended maintaining the F2 generation.

Medaka Multigeneration Test (MMT) and Medaka Reproduction Test (MRT)

Question 1. A rationale for the test method should be available, including a clear statement of scientific basis and the regulatory purpose and need for the test method. The EDSTAC introduced the Tier 2 tests as having the purpose to “characterize the nature, likelihood, and dose-response relationship of endocrine disruption of estrogen, androgen, and thyroid in humans and wildlife.” Tier 2 tests were designed to be definitive tests which generate sufficient data to characterize the specific

hazard of the substance and provide sufficient information on dose-response and adverse effects to permit risk decisions. Please comment on the rationale and purpose of the assay as part of the Tier 2 testing in the EDSP, as described in Sections 2.1 and 2.2 of the MMT ISR.

Panel Response

Item one of the ICCVAM validation criteria document states that “*the scientific and regulatory rationale for the test method, including a clear statement of its proposed use should be available.*” In review of the MMT/MRT ISR (simply MMT ISR), the Panel concluded that each assay provides a clear statement of rationale and purpose. The description in section 2.1 of the MMT ISR is sufficient to justify the need and purpose of the MMT assay, including the necessity for multigenerational testing, full life cycle assessment and appropriate dosing regimen. Overall, the majority of the Panel felt that clear and concise statements were given regarding the advantages of the medaka model, although there was a recommendation to create a specific list of advantages of using medaka as was done in the JQTT ISR. The Panel recommended that more attention be given to the fact that there is a well-defined mechanism to determine both phenotypic gender and genotypic gender in this model. The Panel stated that this aspect of the model is an advantage for utilization in endocrine disruptor chemical studies that is not seen in other small aquarium fish models. The Panel suggested establishing ratios based on phenotypic gender and genotypic gender (e.g., phenotypic to genotypic ratio) within both the MMT and/or the MRT multigenerational assays. As currently written, the MMT/MRT protocols did not take advantage of this aspect of the model. Rather, the sole stated purpose for determining genetic sex is to unequivocally identify gender to establish breeding pairs in the F1 and F2 multigenerational study.

Defined endpoints, including fecundity, fertility, embryonic development, growth and sexual maturity, are easily determined. Overall the Panel found that Section 2.2.2 of the MMT/MRT ISR provides a comprehensive assessment of the “conceptual basis for a multigenerational test paradigm”, but one Panel member thought more detail should be provided for the less informed reader regarding differences between “activational” and “organizational” effects following exposure. The rationale for timing of adult and embryonic exposures was not clearly stated in the MMT ISR (see lines 463-485). In fact, some statements appear speculative such as those addressing chemical permeability across the egg chorion. The Panel did not reach consensus regarding the benefit of the model evaluating maternal-fetal transfer of endocrine disruptor chemicals. One Panel member indicated little information is available on most endocrine disrupting chemicals regarding maternal transfer from the F0 to F1 populations. Another Panel member felt that maternal transfer of many chemicals is a significant concern, as is exposure at various life stages that may vary in sensitivity.

The Panel agreed with the rationale for using medaka as a defined multigenerational endocrine disruption chemical model, i.e., a long history of experimental use, easy to maintain in the laboratory, short generation time, tolerance fluctuations in environmental conditions (salinity, temperature) and can be induced to spawn throughout the year. However, one Panel member felt these very attributes most likely make medaka less representative of native species of concern than the majority of important (ecologically and economically) native fishes. Many native fishes will accumulate chemicals in the eggs (ovary) for months, starting in the fall when they begin to incorporate yolk, until they

spawn in the spring. The young may be exposed to chemicals in the sediment from maternal sources and through the water. In addition, mature fish are likely to be exposed for years, not a few months. For instance, there have been many studies showing endocrine disrupting chemicals adversely affect disease resistance and the immune response and there is no attempt to address this adverse effect in the MMT/MRT assay. Additionally, no assessment of thyroid disruption (e.g., thyroid histology or plasma thyroid hormones) is included in the current MMT/MRT assay protocol. Hence, both the MMT and the MRT predominantly focus on reproductive endpoints and not the other adverse outcomes endocrine disruptors may cause, which can also have population level effects.

The Panel addressed each of the components listed in EDSTAC definition of Tier 2 tests, i.e., “*characterizing the nature, likelihood, and dose-response relationship of endocrine disruption of EAT³ ...*”

- 1) **Nature.** Although non-native and given the weaknesses discussed above, the biological advantages of the medaka as an international model for reproduction should allow the characterization of the nature of most Estrogen, Androgen and Thyroid Hormones (EAT) Adverse Outcome Pathways (AOPs). Its rapid reproduction cycle, secondary sex characteristics (SSC), genetic markers for gender, well-studied developmental and reproductive life history and the ability to evaluate maternal-fetal transport make the medaka an excellent model to characterize the nature of EAT apical endpoints.
- 2) **Likelihood.** For the eight compounds evaluated, there is strong likelihood that E or A⁴ responses may be detected. However, the likelihood of characterizing thyroid effects, or antagonism or indirect effects on steroidogenesis, was less likely. Disruption of the Hypothalamic–Pituitary–Thyroid axis (HPT axis), that is, thyroid homeostasis or thyrotropic feedback control, may influence apical endpoints in this assay, but the response would not likely be specific for thyroid disruption.
- 3) **Dose-Response Relationship.** The data provided significant uncertainty of how dose-response relationships were considered or the meaning of what will be considered significant in relation to surrounding doses if a dose is statistically different from the control, but not from a dose higher or a dose lower than itself. Given the variability of the inter-laboratory calibration studies, particularly those performed by laboratories outside of the EPA research laboratory⁵ located in Duluth, MN (hereafter referred to as EPA MED) and the Japanese National Institute of Environmental Sciences (NIES Japan), obtaining significant concentration-response data outside of single LOEC determinations may be limited in the MMT. If a multiple dose/concentration response curve with more than two statistically significant treatment points (that are different

³ EAT = Estrogen, Androgen, and Thyroid hormones

⁴ EA = Estrogen, Androgen hormones

⁵ EPA, Office of Research and Development, National Human Health and Ecological Research Laboratory, Mid-Continent Ecology Division, Duluth, MN.

from each other) is desired, then increased statistical power or increased range of the dose-response curve is needed. Potential explanations for inter-laboratory variability were provided by the Agency and will be discussed below. Whatever the reason for the variability, most of the MMT data in the ISR demonstrated that single point LOEC measurements had dose/concentration-dependent trends, which limit the diagnostic power associated with the Tier 2 paradigm. If the Agency is only interested in obtaining single point, deterministic measurements of LOEC for apical endpoints that are impaired by EAT AOPs then the MMT is useful. Further, E and A biomarker responses may be linked to the LOEC measurements; however, responses linked to T, EAT antagonism or steroidogenesis AOPs may not be as certain. In addition, monotonic dose-responses are assumed to be dominant in the assays. Overall, the Panel strongly emphasized that caution should be used in evaluating LOECs and other statistical inferences given this assumption since several endocrine disrupting chemicals such as bisphenol A (BPA), may not exhibit typical sigmoidal concentration response relationships (Angle et al., 2013). Enhanced statistical power for the MRT may improve this relationship. However, removal of F0 evaluations for overt toxicity, particularly in females in the MRT, is of concern.

One panelist thought the unique characteristics that make medaka an excellent candidate as a small fish model for this critical bioassay were not described early enough in the MMT ISR. This Panel member suggested inclusion of a clear, bulleted list of the benefits of using medaka as a test animal for a multigenerational assay for testing endocrine disrupting chemicals such as that found in chapter 7 of the JQTT ISR. The Panel developed an example list that includes the following points:

- 1) Medaka are easy to culture,
- 2) Medaka have a relatively small and well annotated genome that is publicly available,
- 3) Medaka are hardy animals and can grow under a broad temperature and salinity range,
- 4) Medaka have anal fin papillae (SSC) that are easily counted and responsive to endocrine disrupting chemical exposures (p.18, MMT ISR),
- 5) Medaka reach reproductive maturation within 2.5 to 3 months, enabling multigenerational assessments,
- 6) Multigeneration studies can be performed in 29 weeks (p. 20, MMT ISR), and
- 7) Medaka have a well described genetic sex determination pathway.

*Question 2. Test methods and their associated endpoint(s) should be scientifically relevant to the biological processes of interest and should be demonstrated to be responsive to the specific type of effect/toxicity of interest. Each species presents unique characteristics from a biological perspective and allows for specialized endpoints to address a specific toxicological mode of action. **Please comment on the biological and toxicological relevance of the assay in regards to the stated purpose of characterizing endocrine disruptors, as described in Section 3 of the MMT.***

Panel Response

In terms of biological and toxicological relevance, relevance was defined by the Agency as “*whether a test is meaningful and useful for a particular purpose*”. Given the unique characteristics of the medaka

life history (as mentioned in response to charge question 1), the biological and toxicological relevance of the assay is theoretically adequate. Specific responses throughout the rapid life cycle allow AOP responses including plasma VTG, SSC, and gonadal histopathology to be evaluated and linked to apical endpoints. The availability of genetic sex markers also allows the evaluation of husbandry conditions which can be refined to provide better reliability and precision.

Numerous endpoints can be considered to assess general toxicity including growth, mortality, hatching success, in addition to defined HPG impacted endpoints. In section 2.2.2 of the MMT ISR, endpoints are described as either “activational” or “organizational”. As such, assessment of SSCs, e.g., formation of papillae on medaka anal fin, can be informative of either androgenic (increase) or antiandrogenic/estrogenic activity (decrease). Assessment of *vtg* mRNA copy number in males is well-known as a “gold standard” of estrogenic effects in oviparous species and is demonstrated to be mediated by ligand dependent transactivation of the estrogen receptor. Thus, VTG is likely to be one of the most sensitive assays proposed for estrogenic effects. The fact that medaka can be induced to spawn daily by controlling light and temperature is also highly relevant to the assay as both fertility and fecundity can be tightly monitored. This component of the assay must be assessed with significant rigor as slight modifications in regulating husbandry conditions may co-impact spawning success in this species. Additionally, there is likely to be significant variability between breeding pairs as demonstrated in the inter-laboratory comparison conducted in this validation process. The Panel recommended that an acclimation period be incorporated into the assay time line to establish baseline data for fecundity and fertility prior to F0 exposures.

An additional advantage is that medaka exhibit phenotypic, sexual dimorphic characteristics relevant to use of this species. However, this aspect of the model is currently underutilized in the proposed MMT/MRT assay. More attention should be given to the fact that there is a well-defined mechanism to determine both phenotypic gender and genotypic gender in this model. This is an advantage for utilization of the medaka model for endocrine disrupting chemical studies that cannot be found in other small aquarium fish models. For this reason, the Panel suggested that both the MMT and MRT assays incorporate genotypic versus phenotypic gender ratios as a defined endpoint of F1 and F2 assessments. In addition, the Panel recommended the development of a population estimate model similar to that of Miller *et al.* (2007) for fathead minnows. An additional concern, particularly with the MRT, is that the assay may not differentiate endocrine responses with chemicals that cause non-endocrine responses. For example, severe hepatotoxicity to the reproductive females in the F0 may alter the endocrine development of F1 offspring providing an “endocrine” response that may not be mediated directly through the E, A or T pathways. While there is adequate redundancy with regard to endocrine endpoints, additional evaluation of hepatic histopathology for severe morphological damage in whole animal cross-sections would allow that particular variable to be assessed for confounding of endocrine responses. Thus, the basic endpoints should be collected in the parental F0 to determine parental toxicity. In addition, the full range of endpoints is needed in the F1, while assessing reproductive function in F1 can be fully assessed by evaluating the F2 through hatch.

The Panel considered it prudent to include within the assay measures of thyroid function. While inherently more difficult than E and A in this model, it is not without precedent that thyroid function is

measurable in these small aquatic vertebrate models. For example, zebrafish measurements have been established for T₄ hormone levels and there is precedent for quantification of transcriptional events in thyroid, brain, liver, and other peripheral tissues that are indicative of thyroid hormone if not HPT function. Bradley et al. (1994) developed an index of HPT function following exposures to prochloraz (PCZ) or propylthiouracil (PTU) using a 20 gene qPCR array in zebrafish. Additionally, there are also previously developed methods for quantifying T₄, including the “immunofluorescence quantitative disruption test” (Raldua & Babin, 2009). The basis of the assay examines the impact of thyroid disrupting compounds to abolished T₄ immunoreactivity in thyroid follicles of zebrafish larvae. It seems this critical function test may be transferable to medaka.

The Panel made several additional comments and recommendations that addressed the biological and toxicological relevance of the MMT/MRT assays in regards to the stated purpose of characterizing endocrine disruptors:

- 1) The Panel recommended that steroidal endpoint measurements (i.e., measurement of estradiol, testosterone, 11-keto-testosterone, and thyroid hormone levels) be included in the protocol to gain further knowledge on the impacts to the endocrine system (see Huggett et al., 2002 and Bulloch et al., 2012, for discussion of methods).
- 2) One Panel member addressed the fact that the MMT does not assess reproductive capacity of exposed males and females independently (e.g., for forming a breeding pair from an exposed male and an unexposed female) and concluded that the MMT is unable to completely resolve impairment of one gender versus the other.
- 3) The Panel commented on the power analysis described in Appendix I, Section 1 (MMT). In this section, the power analysis focuses on fecundity as the key endpoint. If 12 pairs of breeding fish are used then *“the probability of not detecting a reduction of 40% or greater is near zero, and there is probability of greater than 80% of detecting a reduction as low 30%.”* The Panel noted there was no discussion as to whether a median fecundity reduction of 30% is biologically important in medaka or fish in general.
- 4) The power analysis in Appendix I, Section 2 (MMT) focuses on four non-reproduction endpoints, weight, length, SSC and VTG, for 9-week-old medaka. The power analysis for each endpoint is summarized in a table that allows the reader to assess the power of the standard MMT protocol for different fractional reductions as detection differences (10% up to 50%). The reader can look at a range of detection differences before deciding whether the standard MMT protocol is adequate. In addition, a Monte Carlo approach to assessing power (in this case) allows the study designer the opportunity to assess how increasing sample size would improve power for a given biologically important detection difference.
- 5) The proposed MMT has a duration period of 29 weeks, which allows assessment of the adult parental F₀, full life of F₁, and partial F₂ assessment. By comparison, the Panel commented that the MRT study design is more representative of a full life cycle assay with a focus on both

general toxicity and endocrine disruption of a chemical rather than a multigenerational assay. Based on findings from the initial studies, the Panel recommended that the study design for the MRT be changed to end the test after hatching of the F2 populations for a total duration of 19 weeks. Based on the continuous exposure study design a longer F2 assessment does not appear to be needed if the goal is no longer a multigenerational study.

*Question 3. The test protocol should be sufficiently detailed and should include a description of what is measured and how it is measured. The selection of endpoints within the assay should be reflective of the biological processes of interest and the endpoints should be intrinsically relevant and have established sensitivity. The test protocol should demonstrate the ability to measure the endpoints and provide adequate performance criteria for evaluation. **Please comment on the selection, optimization and demonstration of the assay endpoints, as outlined in Section 3 of the MMT ISR.***

Panel Response

The Panel agreed that proposed selection of MMT/MRT endpoints exhibit biological and toxicological relevance with regard to endocrine disrupting chemical exposure and assessment. As demonstrated in section 2.2 of the MMT ISR, small aquarium fish, specifically medaka, present unique biological features that make them highly amenable to *in vivo* endocrine disrupting chemical toxicity testing. Numerous endpoints, e.g., growth, mortality, hatching success, and endpoints with a defined HPG (hypothalamic-pituitary-gonadal) axis are based on established EAT AOPs that provide a foundation for the mode of action of defined endocrine disruptor chemicals. These are described as either “activational” or “organizational” endpoints in section 2.2.2 of the MMT ISR.

- 1) Selection of fish to include in the assay.** Currently there is limited information on how cull choice is made. In section 3.3.3 of the MMT ISR, it is stated that the decision for which surviving larval fish are kept and which are culled at the start of week 4 is not provided. With a maximum of 10 fish per aquaria, some surviving larval fish will be discarded.
- 2) Selection of additional SSCs to evaluate.** Since phenotypic gender alteration is an important endpoint related to endocrine disruption, additional redundancy with secondary sex characteristics may be an option. This is especially of interest given the potential difficulties of *Dmy*⁶ measurements.

In addition to the anal papillae endpoint, Iwamatsu et al. (2003) indicated that additional morphometric characteristics can be recorded for each fish to generate a gender score. The morphometric characteristics included: degree of urogenital protuberance, number of anal fin ray nodes, number of anal fin rays, number of anal fin rays with papillary processes, number of dichotomously branched anal fin rays and degree of ventral fin length. Total length can be determined by a measurement from the most rostral point to the most caudal point of each fish.

⁶ *Dmy* is a Y-specific DM-domain gene required for male development in the medaka fish.

Degree of urogenital protuberance can be quantified on a scale of 1 to 3, with 1 indicating almost no visible protuberance, 2 indicating slight to moderate urogenital protuberance development, and 3 indicating full urogenital protuberance that resembled a half balloon type appearance.

Numbers of anal fin ray nodes can be counted down the third anal fin ray from the fin origin. Numbers of anal fin rays with papillary processes can be counted for each ray with any visible papillary processes. Degree of ventral fin length can be quantified on a scale of 1 to 3 with 1 indicating a short ventral fin not extending to the genital pore region, 2 indicating a ventral fin that extended to the genital pore region, and 3 indicating a ventral fin extending to or beyond the genital pore.

A gender determination score can then be generated for each fish in control and treatment groups. Score can be generated by adding degree of urogenital protuberance index, number of dichotomously branched anal fin rays, and ventral fin length index for each fish. Male fish have only one or two dichotomously branched anal fin rays, whereas female fish have six to eight. A greater gender determination score might be attributed to an individual fish exhibiting female morphometric characteristics and a lower score for fish with more male characteristics. Automated programs are available for these types of morphometric analyses.

3) Optimization and demonstration of assay endpoints. Optimization of MMT endpoints appears to be critical to establishing rigorous statistical inference. As such, the MRT assay is a MMT assay that has been modified to reflect several changes including increase in biological replication for fertility and fecundity assessment. Other endpoints, such as SSC, appear to be highly correlative to endocrine disrupting chemical AOPs, MOA and indicative of endocrine disrupting chemical function. Overall, a relatively small number of compounds with differing endocrine disruption MOAs have been studied using the MMT/MRT assay and thus, additional chemicals should be tested before this assay is considered validated. To ensure consistency in evaluation of individual defined endpoints, further standardization and possible optimization of methodology may be required. As is, section 3 of the MMT ISR provides a reasonable global description of the MMT/MRT assay including timelines, exposure conditions, husbandry, and overall measurement of defined AOP and general toxicity endpoints. However, the MMT ISR and provided protocols, lack sufficient methodology to fully reproduce this assay. This is particularly true with regards to methodology and data analysis for each defined endpoint of the assay. For example, the following elements should be included in a standardized protocol.

- **Water conditions.** While water conditions were briefly mentioned, water quality is a major differential factor among laboratories and more details should be provided in the research protocol on water quality requirements. This should also help to reduce variability among laboratories that will be conducting these studies. Currently, limited information is provided for additional water quality parameters (e.g., hardness, salinity/conductivity, pH, etc.). Most small fish laboratories have extensive protocols developed for water quality guidelines and similar guidelines should be included in this assay to increase its optimization.

- **Reproductive activity of fish mating pairs.** Assessment of the reproductive activity of the fish mating pairs prior to F0 exposures, and detailed husbandry guidelines should be added that indicate that the size of the adults and reproductive capability of that particular fish population should be assessed before starting this assay. This is important as these basic population characteristics can vary widely among different laboratories. Adding standardization on minimal size and reproductive function of the specific fish population to be used in the studies should help assist in removing some of the variability among the different laboratories. In addition, most small fish laboratories also have extensive protocols developed for husbandry and care of the fish. Similar guidelines should be included in the protocol for guidance and optimization of the assay across laboratories.
- **Standardization of VTG analysis.** Additional descriptions need to be added for this analysis.
- **Analysis of anal fin papillae.** Overall a detailed description of the anal fin papillae should be provided with some description on counting papillae structure.
- **Embryo collection standardization.** Additional guidelines need to be added for this procedure.
- **Embryo viability.** More details on how viability is determined are needed.
- **Determination of genetic sex.** Additional descriptions need to be added for this analysis. As an example, the inter-laboratory analysis was complicated by the fact that there was supposed to be contamination of this assay component. This resulted in conflicting data between laboratories with an apparent increase in data variability.
- **Hepatotoxicity assays.** The type of evaluations and endpoints studied in hepatotoxicity assays need to be clarified. It is unclear what types of toxic responses are expected or how histological findings will inform endocrine disrupting effects of test compounds. It is important to establish normalization procedures for histopathology assays to insure assay reproducibility.

More information on the specific details on what to include for these endpoints can be found in the response to question 6. Overall, these additional details and parameters will assist in optimizing the assay.

Question 4. Demonstration of the test method performance should be based on the testing of reference chemicals representative of the types of substances for which the test method will be used. Test substances should adequately represent an appropriate range of responses and physical/chemical properties for which the test method is proposed to be appropriate. The selection of the most appropriate statistical approaches depends in part on the nature of the data and also on the design of the validation study. Statistical and non-statistical methods used to analyze should be described.

Please comment on the selection of test substances and methods (analytical and statistical where appropriate) chosen for the demonstration and validation of the MMT and MRT assays.

Panel Response

1) Selection of test compounds

The following test compounds were used to validate the MMT assay:

- The four ER agonists used were: 17 β - estradiol (listed as strong), *o,p'*-DDT (medium), 4-*t*-octylphenol (weak), and 4-chloro-3-methylphenol (weak).
- Tamoxifen is listed as an ER antagonist and a weak ER agonist.
- Trenbolone is listed as a strong AR agonist.
- Prochloraz is listed in the MMT ISR as an aromatase inhibitor with androgen receptor antagonist activity. However, there is at least one previous study with amphibians that shows prochloraz can also disrupt the thyroid endocrine axis and delay metamorphosis (Brande-Lavridsen *et al.*, 2010). The Panel recommended the Agency add that prochloraz has a thyroid-dependent AOP.
- Vinclozolin is listed as an AR antagonist and is used in the ring test.

Vinclozolin was used in the MMT on the basis of a presumptive AOP that is initiated by binding to the AR without activating androgen-dependent genes. However, a recent study with rats presented evidence suggesting that vinclozolin is also a thyroid-active compound that affects thyroid hormone production and/or metabolism (Schneider *et al.*, 2011), and another study using human recombinant thyroid peroxidase (TPO) showed that vinclozolin is able to suppress TPO activity (Song *et al.*, 2011). Although there is no information concerning the ability of vinclozolin to act as a thyroid endocrine disruptor in fishes, the Agency is encouraged to review the data for vinclozolin that was presented in the MMT ISR in the context of potential effects on both the androgen and thyroid endocrine systems. Also, given the reported association between disruption of the thyroid endocrine axis and unbalanced sex ratios in some species of teleosts and amphibians (Carr & Patiño, 2011), it was notable that the Agency's vinclozolin study reported mismatches between genotypic sex and gonadal sex in some of their fish treated with vinclozolin (pp. 35-36 of the MMT ISR).

2) Selection of statistical approaches

The Panel summarized the MMT test conditions for the EPA MED⁷ and NIES Japan⁸ studies described in the MMT ISR (Table 3.1). Briefly, the MMT protocol consists of continuously exposing 2½ generations of fish to the test chemical starting with maturing adults in the first generation (F0) and

⁷ EPA MED = EPA, Office of Research and Development, Mid-Continent Ecology Division, National Health and Environmental Effects Research Laboratory, Duluth, MN

⁸ NIES Japan = Japan, National Institute of Environmental Sciences

exposing them for several weeks to allow chemical accumulation. There are six treatments (if a solvent is not used); five chemical treatments and an unexposed control treatment. There are six replicates aquaria in each treatment.” The Panel referred to “aquaria” as “tanks” in these comments.

The Panel stated that Table 3.1 in MMT ISR does a very good job of describing those conditions that can be controlled by the researcher, but a poor job in describing randomization of the fish to these conditions. The Panel raised several questions: How many flow-through systems are there? Is there one test replicate chamber size 18x9x15 cm in each flow-through system? Or are all test chambers placed in the same flow-through system? The Panel thought there was a single flow-through system for each tank. Is this true? When is feed placed in a tank? Is this placement done in a random order? Where are the tanks placed? Are their locations randomized? The Panel stated that many details are missing on how to handle the tanks at each generation. Without these details, anyone who tried reproducing the same conditions would have great difficulty. The Panel recommended that the experimental conditions be described much more explicitly.

In Table 3, only two adult control parameters, fecundity and SSC, in two octylphenol MMTs were comparable between two laboratories, EPA MED and NIES Japan. The Panel did not understand why these data were similar given that the means for fecundity and SSC in the EPA and Japan studies are significantly different given the standard deviations shown.

Table 3. Comparison of adult control parameters in two 4-*t*-octylphenol MMTs

	US EPA MED		NIES Japan	
	F1	F2	F1	F2
Fecundity	17 ± 2	20 ± 5	29 ± 5	27 ± 3
SSC	83 ± 9	130 ± 5	116 ± 12	120 ± 19

Source: Table 4-3, MMT ISR

Originally, the Panel had some additional questions about the EPA’s corrected statistical analyses of the MMT data. The Panel explained their confusion was the use of “repeated measures analysis” in the EPA’s explanation of the corrected statistical analysis. During the meeting, the EPA indicated that “repeated” was a misnomer, as the program actually used a mixed model analysis with tank as a random effect. The Panel agreed that this was the correct way to analyze the data with unequal numbers of observations per tank.

*Question 5. Considering the variability inherent in biological and chemical test methods, a test method needs to be repeatable and reproducible. A test is robust and reliable if the results are repeatable and reproducible within a laboratory and between different laboratories, respectively. A test protocol should provide sufficient guidance to ensure proper and consistent performance across labs and chemicals. **Please comment on the test method robustness and reliability and the repeatability and reproducibility of the results obtained with the MTT and MRT assays.***

Panel Response

For the purposes of this question and consistency in the discussion across assays the Panel defined the following terms, “reliability” and “robustness”. Reliability is defined by ICCVAM as “the reproducibility of results from an assay within and between laboratories.” Robustness is defined as a “characteristic describing a model's, test's or system's ability to effectively perform while its variables or assumptions are altered. A robust method is one that can operate without failure under a variety of conditions.” In general, being robust means a system can handle variability and remain effective (referenced at <http://www.investopedia.com/terms/r/robust.asp>). Robustness is also used to refer to the performance of the assay in relatively “inexperienced” laboratories.

Each of the reference chemicals used for assay validation was associated with one or more endocrine-dependent AOPs that were summarized in the MMT ISR. At a basic level, the general reliability of the assay can be evaluated initially by comparing effects of each test chemical against expected endpoint results of their respective AOPs. The MMT assay generally performed as expected for the reference materials. The Panel noted that both EPA MED and Japan NIEHS are experienced with the test chemicals and have extensive experience in conducting these assays; therefore, a high level of assay reproducibility can be achieved. Less experienced laboratories with conducting the MMT/MRT may not achieve a high level of assay reproducibility.

- 1) All ER agonists typically resulted in reduced fecundity and, in males, decreased secondary sex characteristics (anal fin papillae), up regulation of hepatic *vtg* mRNA, and feminization of gonads.
- 2) Tamoxifen, an ER antagonist with some agonist activity, also reduced fecundity. In addition, in females tamoxifen also reduced *vtg* mRNA, an expected result in view of its ER antagonist activity; and in males, tamoxifen reduced SSC and increased *vtg* mRNA, also as expected given its ER agonist activity.
- 3) The androgen agonist resulted in decreased fecundity and, in females, also caused development of male-specific SSC and reduced hepatic *vtg* mRNA (likely via disruption of HPG axis) as well as sex-reversal (masculinization) of XX (*Dmy*-negative) fish.
- 4) Unlike the situation for most other teleosts (Guiguen *et al.*, 2010), aromatase inhibitors do not result in sex-reversal of genetic female medaka, but they do disrupt their normal ovarian development (Suzuki *et al.*, 2004). Therefore, exposure to aromatase inhibitors is expected to lead to reproductive impairments. In the MMT assay, the aromatase inhibitor, prochloraz caused a reduction in fecundity, inhibition of hepatic *vtg* mRNA in females, and had no clear or consistent effects on SSC in males. These observations are all generally consistent with the anticipated effects of an aromatase inhibitor in medaka.
- 5) Because the reported effects of various androgen antagonists have been inconsistent or ambiguous in the literature, the ISR did not specify expected results for vinclozolin and proper assessment of its performance cannot be made based on the information provided. As already noted, however, vinclozolin has also been shown to be a thyroid-active compound (see response to MMT charge question 4).

The MMT ISR presented two sets of inter-laboratory assays to focus on reproducibility between laboratories. In the first set, the EPA MED and NIES Japan independently examined the effects of 4-*t*-octylphenol (weak ER agonist) using the full MMT. Although the test format was not identical in both studies, the differences were small enough to allow for valid comparisons. Results between the two tests were generally consistent; both reported impaired reproduction (reduced fecundity); reduced SSC, and induction of hepatic VTG in males; and sex-reversal or inter-sex induction in XY individuals. For fecundity, SSC, and hepatic VTG, LOEC values between studies were also in good agreement. These observations suggest that when using a test chemical with a fairly well defined AOP by well-established laboratories with extensive experience in the test protocols, a high level of assay reproducibility can be achieved.

The second round of assays involving three contract laboratories, which was the formal ring test for the MMT, used vinclozolin as the test chemical and an abbreviated MMT format (24 weeks instead of 29 weeks in the full MMT). The results of these assays were not encouraging. Several major inconsistencies or technical difficulties were noted as follows:

- 1) The measured concentrations of vinclozolin were apparently difficult to achieve and maintain in two of the three laboratories.
- 2) Several of the reproductive endpoints in the control F1, as well as growth endpoints in the control F1 and F2, were significantly different among the laboratories. The ISR attributed this variability to the unusually small size achieved by fish in one of the laboratories (Laboratory 2) and, in fact, correcting for size eliminated the inter-laboratory differences in fecundity for control animals.
- 3) Because there were problems with the measurement of SSC in one laboratory (Laboratory 3), three-way comparisons for F1 were not possible for this endpoint. Also, although the ISR states that when data were available the findings were consistent for this endpoint, this appears not to be the case. Namely, unlike Laboratory 1, Laboratory 2 did not observe a change in SSC in F1 males.
- 4) The EPA MED vinclozolin, full MMT study, as well as those of the contract laboratories using the MRT, had a problem in that a number of their fish showed genotypic sex that did not match their gonadal (histological) sex. This problem was much more serious in assays conducted by the contract laboratories, which reported high levels of both XX-male and XY-female mismatches. The ISR attributed this problem to potential artifacts of contamination during sample collection or processing. This is probably a correct assessment in the case of the contract laboratories because they also reported high levels of mismatch in their control fish, whereas, the Agency study did not. However, because vinclozolin may have alternative AOPs which potentially could cause sex-reversal in fish (see answer to MMT charge question 4), at least some of the mismatches may have been true biological responses to vinclozolin.

Overall, the Panel stated that an evaluation of reproducibility for the ring test using vinclozolin is difficult because of: 1) the relatively undefined AOP of the test chemical used (vinclozolin) and, therefore, of the expected endpoint results, 2) the technical problems encountered by the performing laboratories, and 3) the largely negative results. To possibly shed some light on reasons for the relatively

poor performance of the ring test, the Panel recommended that the Agency consider, at least conceptually, the possibility of alternative AOPs for this test compound.

In conclusion, while the inter-laboratory study between EPA MED and NIES Japan with 4-*t*-octylphenol showed similar data responses between laboratories, the vinclozolin results were much more inconsistent. The EPA MED and NIES Japan laboratories have considerable experience with the husbandry of the medaka model, and consequently, less variability is expected from studies conducted in these laboratories. However, the ring test with vinclozolin is likely more relevant to laboratories that would be conducting the Tier 2 tests in the future. Therefore, for practical purposes, the Panel believed that reliability and robustness has not been fully demonstrated for the MMT.

To improve reliability and robustness of the assay, the Panel recommended that the Agency develop more specific and detailed standard operating procedures and guidelines to conduct the assay (see answer to MMT Question 6). Greater emphasis needs to be placed on the overall health of fish populations being used in the studies. In particular, procedures and conditions for the adequate care and husbandry (including specific ranges for dilution water quality variables, such as hardness and micronutrients, need to be better highlighted in the guidance.

Question 6. The test protocol should be descriptive enough to be fully transferable to a functional laboratory. The protocol should describe the methodology of the assay in a clear and concise manner so that a laboratory could comprehend the objective, conduct the assay, observe and measure prescribed endpoints, compile and prepare data for statistical analyses, and report the results. Section 4 of the MMT ISR outlines the process of and challenges experienced in the inter-laboratory studies and Section 5 presents the optimization and proposed Medaka Reproduction Test (MRT) protocol. Please comment on the transferability across labs and provide any suggestions or recommendations for improvement of the MMT and MRT assays.

Panel Response

To evaluate the transferability of the MMT assay, the Panel referred to the ICCVAM validation criteria which stated that, “*within test, intra-laboratory and inter-laboratory variability and how these parameters vary with time should be evaluated.*” According to the MMT ISR, there were three data sets that could be evaluated for intra-laboratory and inter-laboratory transferability. These include a set of inter-laboratory MMT trials with 4-*t*-octylphenol, assessment of the control values from the same study and a vinclozolin ring test.

Data presented in Table 5-7 of the MMT ISR suggest there is modest agreement across trials for a subset of parameters in the inter-laboratory validation studies with 4-*t*-octylphenol (see also discussion and recommendations in response to charge question 5). For example, the EPA MED 2007 trial and the NIES Japan trial reported similar fecundity LOEC values for 4-*t*-octylphenol when a common assay protocol (with slight modifications between laboratories) was followed. In 2012, EPA MED refined the MMT, including more replicates, and reanalyzed the effects of 4-*t*-octylphenol, which produced significantly different results.

The strongest argument in support of transferability is made by assessing control parameters across MMT trials. Section 4.1.4 (MMT ISR) illustrates such a comparison between MMT trials with 4-*t*-octylphenol. Here, the ISR describes that the MMT trials between EPA MED and NIES Japan exhibit significant consistency/reproducibility for fecundity and SSC in both F1 and F2 populations. This section continues to discuss successful reproducibility in the comparison of additional endpoint LOEC data for 4-*t*-octylphenol. However, the Panel noted that the EPA MED vs. NIES Japan data comparison likely demonstrates reproducibility rather than transferability because the NIES Japan laboratory trial was conducted in a laboratory environment highly experienced with medaka techniques and had existing protocols that only needed slight modifications to comply with the MMT assay (i.e., this laboratory designed the test and was the most experienced in conducting the test). However, the Panel indicated that the MMT protocol used by NIES Japan was executed somewhat differently than by EPA MED suggesting that further refinement and clarification should be included to the MMT to ensure global adherence to the same MMT protocol and QA/QC parameters

One inter-laboratory ring test was conducted between three contract facilities to assess transferability of the MMT. The ring test examined MMT assay performance when challenged with AR antagonism as a defined AOP using vinclozolin as the chemical test agent. Results from the inter-laboratory ring test were highly variable and suffered from a lack of consistency in both execution of the MMT protocol and overall assay performance. The Panel also raised significant concern regarding the choice of both AOP and chemical agent tested for the initial ring test because the toxicological mechanism of vinclozolin and its metabolites can be complicated and both biochemical and apical endpoints for AR antagonism in medaka are not well defined. As presented, results from the vinclozolin study clearly indicate that the MMT assay, and likely MRT assay, is not currently transferable to other laboratories.

As stated in Appendix A of the MMT ISR, the ICCVAM validation guidelines require that, “*A formal detailed protocol must be provided and must be available in the public domain. It should be sufficiently detailed to enable the user to adhere to it and should include data analysis and decision criteria.*” Given this guidance criteria, the Panel suggested that both the MMT and MRT protocols include a refined methods section that specifically detail all components of the assay as designed. As presented in the ISR, there is an inadequate description of the assay protocol and defined methodologies of general experimental principles including: experimental design, biological/technical replicates, husbandry/culture conditions for assay performance, test agents used and concentration range, definable primary and secondary endpoints of the assay, water chemistry, and overall timing of multigenerational approach.

There are noticeable components of the MMT and MRT assays that have no documentation that would significantly aid in reducing data variability and enhance overall concordance and transferability of the assay within inter- and/or intra-laboratory trials. While it is expected that most laboratories will have a working knowledge of medaka biology, the Panel recommended detailed protocols be developed to provide sufficient (detailed) guidance for all measurable endpoints.

As an example, detailed guidance for measurement of *vtg* gene expression by qPCR should include information regarding:

- 1) Tissue (liver) collection. How is liver dissected, what washes are used, how is liver stored, for how long, under what conditions, is an RNA stabilizing solution used or is tissue flash frozen?
- 2) Standardize RNA extraction reagents and methods. What reagents are used for RNA extraction? How is RNA QA/QC, concentration and quality determined?
- 3) cDNA synthesis amplification reagents and methods. What reagents are used for cDNA synthesis, random priming or poly A-oligos? What is the concentration of RNA used in the reaction, what is the RT (reverse transcriptase) incubation time and what is the type of RT used?
- 4) qPCR primers. A standard set of medaka *vtg* primers should be listed that target *vtg I* or *vtg II*. Which *vtg* is targeted for the assay, *vtg I* or *vtg II*, or both? Which reference gene was evaluated and the primers for this gene?
- 5) qPCR assay conditions. All amplification conditions should be noted, i.e., dissociation temperature and time, annealing temperature and time, extension temperature and time, number of cycles, concentration of dNTPs, buffer and technical replicates.
- 6) qPCR analysis. Data analysis and numerical process of determining *vtg* copy number should be described.
- 7) Reference standards for *vtg* copy number (standard curve) should be developed for quality control between laboratories.
- 8) Overall guidelines should make sure that MIQE: Minimum Information for Publication of Quantitative Real-Time PCR guidelines are being followed for this analysis (see Bustin et al., 2009).

There were a number of additional measurable endpoints that the Panel stated needed further QA/QC and performance guidance, including:

- 1) Analysis of anal fin papillae. Overall detailed description of the anal fin papillae should be provided with some description on counting papillae structure. It may be prudent to count papillae within defined parameters such as omitting most proximal strands (see the response to charge question 4).
- 2) Embryo collection. There are several methods of embryo collection and removal of embryonic threads, i.e., manual removal via scissors and tweezers, rolling embryos on fine grit sandpaper, etc.
- 3) Embryo viability. How is viability determined? Is methylene blue used or is viability assessed by opaqueness of the embryo, timing of assessment, i.e., number of hours post fertilization.
- 4) Determination of genetic sex. As per *vtg* analysis, all methods regarding sample collection and storage conditions, DNA extraction and qPCR conditions should be included as an appendix to the assay. As an example, the inter-laboratory analysis was complicated by the fact that there was purported contamination of this component of the assay. This resulted in conflicting data between laboratories with an apparent increase in data variability. *Dmy* reference material would help alleviate this issue.
- 5) Water quality parameters were not well described. Given the significant differences between water sources, additional detail should be provided, particularly for hardness and micronutrients such as Mb and Se.

- 6) Reference materials for *Dmy* and *vtg* endpoints through the National Institutes of Health should be developed to assess assay performance.
- 7) Fish husbandry and care. Most fish laboratories have very detailed lengthy protocols for daily and routine care procedures. These protocols generally include water quality source, water quality parameters and feeding parameters, including artemia hatching, etc.
- 8) Fish health assessment. Overall, laboratories should have a good grasp on the health and activity of their fish population prior to initiating MMT/MRT assay to ensure that population health and reproduction is adequate.

Consequently, the Panel suggested the protocol be re-evaluated to determine if clarity can be improved. It is recommended that both MMT and MRT incorporate reference toxicants and materials that can be used for internal assessment of responses. Reference toxicants that encompass E, A, AE, AA or T should be used if disruption of these pathways is identified in Tier 1. Thus it is the opinion of the MMT/MRT review that an inter-laboratory comparison should first be examined with a reference ER agonist/antagonist, a reference AR agonists/antagonist and a reference TR agonist/antagonist (if included in the assay) with well-defined chemical and biological behavior. Also unresolved is whether a data manipulation strategy is justified to mitigate assays that do not perform well. More clarity is needed here. Lastly, the justification for a non-solvent control does not seem to be clear in the documentation. The Panel indicated that laboratories, implementing the recommended refinements and strict adherence to designated MMT and MRT assays protocols, will likely perform with a degree of repeatability and reliability. Subsequent assessments could then be made with chemical agents that test less definitive AOPs (such as AR antagonism) to assess apparent sensitivity of assay conditions across laboratories. Within this context, additional inter-laboratory studies (outside of EPA MED and NIES Japan) are needed with specific guidelines to prove transferability can occur.

Question 7. The purpose of the validation process is to determine the readiness of a test for inclusion in a testing program. A component of readiness of a test is the evaluation of the usefulness and limitations of the test, including the classes and types of test substances that can and cannot be tested. Please comment on the strengths and/or limitations of the MMT and MRT assays.

Panel Response

Strengths

Overall the Panel identified numerous strengths of the MMT and MRT assay that related to both the use of medaka as a model organism and experimental design and assay protocol.

1) Model organism

Advantages of using a small aquarium fish model:

- Husbandry conditions well established
- Ease of manipulation
- Well sequenced genome for reference

- Rich history with available resources in endocrine toxicity
- Short generation time to decrease assay time
- Routine and external spawning with daily spawn
- External embryonic development
- Transparency of the embryos (enables possibility to assess early life stage toxicity, for general toxicity assessments, if desired).

Ability to test multiple endocrine disrupting chemical AOPs via:

- Genotypic sex determination
- Phenotypic sex determination
- Well established and reliable biomarkers of endocrine disrupting chemical effect/exposure (i.e., VTG and SSC)
- Statistically robust (i.e., use of large numbers of animals/replicates)
- Ability to perform multiple generation assays

2) Experimental design and assay protocol

- The MMT assay includes assessment of F0 through F2 to reproduction. The full assessment of reproduction through two generations allows the evaluation of epigenetic effects that may occur as a result of EAT disruption during *in ovo* exposure or development.
- Strengths of usefulness of this assay include inclusion of chemical exposure in F0 to assess parental transfer, organization and activation effects.
- Strengths specifically related to the MRT assay include a reduction in assay time, improved power of the fecundity/fertility test due to the replicate nature of the experimental design.

Limitations

The Panel identified numerous limitations of the MMT and MRT assays that were specifically related to experimental design and clarity of the assay protocol.

- 1) MMT has lower power than the MRT assay for statistical comparisons (important given limited “causal” dose/concentration dependence in documented studies).
- 2) The MMT assay has higher labor requirements, time, animal use and overall cost than the MRT assay.
- 3) There are a limited number of MMT trials with representative chemicals associated with each endocrine disrupting chemical AOP. Additionally, there is a need to include reference chemical toxicant that is not mediated by endocrine disrupting chemical effects, such as a reference hepatotoxicant like α -naphthylisothiocyanate.
- 4) There is less certainty with AOPs that represent anti-estrogens and anti-androgens.

- 5) There is a lack of standardized protocols within the assay for measurement of individual biochemical and apical endpoints.
- 6) Reproductive endpoints may be sensitive to animal husbandry conditions.
- 7) No information was provided on medaka strains used; thus, chemical sensitivities may vary between laboratories and/or countries.
- 8) Inter-laboratory reproducibility is yet unproven.
- 9) There is a need for greater replication in MMT, an advantage for using the MRT.
- 10) MMT/MRT can be informative of endocrine disrupting chemical AOP, but may be complicated by compounds with putative multiple responses (e.g., tamoxifen can be an ER agonist/antagonist or prochloraz, which may function as aromatase inhibitor and or AR antagonist). Thus, endpoints will not definitively inform the mechanism.
- 11) The full potential of phenotypic sex is not utilized, i.e., ratios of phenotypic sex to genotypic sex can be determined.
- 12) There is a lack of proper training and guidance to laboratories to perform the assay.
- 13) There is no assessment of animal health prior to study.
- 14) Statistical power of MMT assay is improved with MRT; however, this is at a cost of reduced assessment in F0 and F2 generations.
- 15) In MRT, loss of F0 and F2 endpoints may result in putative loss of essential data for AOP and toxicity inference.
- 16) Data related to the effects of steroidogenic disrupting compounds were not as thoroughly evaluated over multiple generations. Consequently, it may be premature to conclude that F2 generation studies are not necessary. Studies including steroidogenesis disrupting compounds add a greater diversity of MOAs evaluated over more than one generation would improve confidence in the assay.
- 17) There is a need for correlation/integration between Tier 1 fathead minnow studies and the Tier 2 MMT/MRT.

*Question 8. There is sufficient evidence to indicate that endocrine disrupting chemicals can disrupt normal development and reproductive success, however the value added of the F2 generation compared to the P0 or F1 is less clearly defined. The EPA is proposing the Medaka Reproduction Test (MRT) as an EDSP Tier 2 assay. The MRT test terminates after the F2 embryo hatch, and a rationale for the proposed protocol is provided in Section 5 of the MMT ISR. **Please comment on the Agency's rationale that the value added by the F2 generation is not sufficient to warrant its inclusion in the Tier 2 fish test protocol.***

Panel Response

The current structure of the MMT assay represents a true multigenerational assay with reproductive development of F1 and F2 fish examined. Removal of the F2 generation in the MRT then raises the question of whether the MRT is a truly “multigenerational” test. As designed, chemical exposure in the F0 could impact F1 embryos by maternal transfer or modulation of maternal factors. Information regarding activational and organizational effects of endocrine disrupting chemicals may be lost (e.g.,

effects on mating behavior). This is particularly true for inductive responses observed in the F0 adults, such as *vtg* expression.

Comparison of data across MMT trials suggests that analysis of biomarker and histopathology data for the F2 generation may be redundant as no significant difference in these endpoints was observed between the F1 and F2 generations. For example, Tables 5-7 and 5-8 (MMT ISR) indicate concordance between F1 and F2 LOEC values for fecundity and biomarker data, respectively. These data and additional discussion in the MMT ISR provide a strong case for removing the F2 reproduction portion of the MMT. Rationale for the MRT is that it will increase statistical power, reduce animal number, reduce person hours to perform the assay, reduce overall assay time and reduce possible variability of data due to increased risk of experimental incidents resulting in test failure. Overall, this rationale indicates that data from an additional generational would not increase the quality of the risk assessment.

Assessment of the F1 generation is likely to inform both activational and organizational responses and will link the responses to endocrine disrupting chemical toxicity in a single generation. However, exclusion of the F2 effects will limit the MRT to a single generation assay or at most a 1.5 generation assay. The Panel suggested the MRT assay could be modified to allow F2 animals to reach sexual maturity (~9 wpf⁹) to derive additional data regarding both phenotypic and genotypic gender development. As currently described in the MMT ISR, neither the MMT assay nor the MRT assay are able to establish genotypic to phenotypic gender ratios. Establishing this analysis (with data already derived from existing assay structure) will further substantiate modifications in phenotypic gender as the genotypic sex of the individual fish will not change with chemical exposure.

Given the current data set, the Panel concluded that termination of the medaka assay at F2 hatch appears warranted without significant loss of sensitivity for endocrine endpoints. However, this conclusion has the following caveat, the data to support the removal of the F2 were from minimal assay trials that had only one comparison of MMT with MRT using 4-*t*-octylphenol. Animal number and labor costs will also be minimized. However, it is recommended that not all F0 endpoints be discarded. In particular, endpoints that evaluate organ histopathology or overt toxicity (growth) should not be removed due to the potential confounding of non-endocrine maternal effects. The Panel noted that the data to support the removal of the F2 were from minimal assay trials that had only one comparison of MMT with MRT using 4-*t*-octylphenol. With the continuous exposure design in the MRT assay, the Panel expected that results will be similar in F1 and F2. The Panel suggested the addition of tissue chemical analysis to aid in determining if bioaccumulation of the chemical in the tissue is occurring. This would lead to an increase in the overall number of animals for the study.

While compounds tested represent a range of endocrine responses, there was still significant uncertainty with AR antagonism and steroidogenesis AOPs. It is unclear, based upon the current data, whether a more potent AR antagonist or aromatase inhibitor might affect F2 endocrinology. One panelist felt that the answer to this question depended on a clear statement of the key endpoints for each

⁹ wpf = weeks post-fertilization

endocrine disrupting chemical study, and this would depend on the hypothesized AOP. For example, if the endpoint is body weight and the observation is increasing body weight in subsequent generations, this might indicate an endocrine-related effect on obesity. If this is the case then it would be beneficial to collect F2 adult body weights. However, the Panel noted that body weight, other than as a measure of vigor and readiness to reproduce is not typically considered a reproductive endpoint of concern. The focus of the MMT is on reproductive endpoints.

Question 9. It is the Agency's opinion that the outcomes of the various MMT trials have provided enough information to recommend a medaka reproduction test (MRT) for use as the fish test in Tier 2 of the EDSP. Two major changes from the MMT are proposed, i.e., an increase in the number of replicates per treatment for evaluating effects on reproduction, and terminating the test after the embryos hatch in F2. Other proposed changes include minimizing the collection of endpoint data from F0, and evaluating pathology in only the F1 adults sampled after the assessment of reproduction.

*Overall, the authors conclude that both the MMT and the MRT are transferable methods and are capable of adequately characterizing potential disruption of the endocrine system by putative endocrine disrupting chemicals. However, the MRT is recommended as the preferred EDSP Tier 2 test method for fish because it is less resource intensive with improved statistical power, appears to be as sensitive, and is better able to ensure consistent findings when performed routinely by testing laboratories. **Please comment on the scientific rationale of the Agency's proposed Tier 2 fish test, the MRT, with respect to statistical power, sensitivity, and consistency in performance across laboratories. In addition, please comment on the adequacy of the MRT to characterize potential endocrine disruption, a requirement of Tier 2 of the EDSP.***

Panel Response

The majority of the ISR focuses on MMT with Section 5 providing a rationale for the MRT. The MRT differs from the MMT in two fundamental ways: 1) increased replication of breeding pairs for evaluations of reproduction in the F1 generation and 2) termination of the test after hatching of F2 embryos with no biochemical assessment of the F2 individuals. Based on statistical considerations, evaluations of reproductive parameters will be much improved with increased biological replication and are likely to enhance accuracy of acquiring precise LOEC data from reproduction endpoints in the F1 population. Detailed modeling of reproductive replicates with Monte Carlo simulations strongly supports this modification and trial experiments with 4-*t*-octylphenol demonstrated increased sensitivity, enabling detection of a lower reproductive LOEC (23 µg/L) compared to the MMT (100 µg/L) trials. However, in the MMT, the LOECs based on biomarker assays (many of which will not be completed for more than one life stage or generation in the MRT) were lower, 13 µg/L and 6.25 µg/L, respectively (depending on the laboratory), than the reproductive endpoints in the MRT.

The MRT assay has many fewer endpoints than the MMT and, hence, there is some concern about the impact of modifications to assess multigenerational effects as well as to assess non-reproductive parameters. In the MRT, no assessments are made on F0 animals; *vtg* expression is only conducted in sub-adults at 9 wpf, and histopathology only in F1 adults at week 15 after spawning. While the rationale for these modifications is given, exclusion of some of these endpoints may reduce the overall assay

performance. For instance, histopathology analyses should be included at 9 wpf as there is significant variability in growth and development of medaka at this stage. Given that *vtg* expression will be assessed at only this stage, some supportive assessment of gonad development and maturation should be determined. A similar argument can be made for determinations of *vtg* expression in the adult F1 generation as these animals are expected to have fully mature gonads. It is unknown if correlations can be demonstrated among *vtg* mRNA levels, gonadal morphology, degree of intersex or other gonadal abnormalities and fecundity. Hence, important data may be lost by only having one of these endpoints in a particular generation. The Panel recommended further effort should be made to coordinate all three endpoints. Additionally, F0 animals should be evaluated for histopathology to determine whether non-endocrine effects or overt toxicity (particularly at the maternal stage) confound the assay and indicate “endocrine” AOPs in offspring. For example, overt toxicity may diminish blood flow to ovaries or cause liver damage which may lead to responses in reproduction that are not mediated through EAT AOPs. Along these lines, non-endocrine toxicants that may indirectly affect reproduction should be evaluated during validation of the MRT assay.

The Panel recommended that additional confirmatory studies under more specific and detailed guidelines be performed before MRT is adopted as a Tier 2 assay. Overall, it appears that global validation of the MMT/MRT is rather subjective. A process of producing data consistent with ICCVAM validation criteria has been established. However on the global scale, there is little statistical indication of performance parameters needed to unequivocally establish assay reproducibility and reliability. Additionally, inter-laboratory studies are needed to demonstrate transferability, reliability and reproducibility.

In conclusion, the Panel stated that the MRT showed clear promise with regard to statistical power, sensitivity and consistency in performance with the 4-*t*-octylphenol test, and the rationale for modification is adequate. However, given the uncertainties inherent in fish husbandry and endocrine variability with this species, additional confirmation with other chemicals is necessary for validation. The 1998 EDSTAC recommended 50-100 compounds go through Tier 1, but it was unclear how many should go through Tier 2. The Panel recommended that representative compounds (both low and high potency) from the ER, ER-A, AR, AR-A, T, T-A and steroidogenesis inhibitor AOPs, as well as a nonendocrine AOP, be evaluated for MRT assay performance prior to adoption as a Tier 2 assay. Ring testing with the MRT (and perhaps one of each AOP) is also recommended.

Larval Amphibian Growth and Development Assay (LAGDA)

Question 1. A rationale for the test method should be available, including a clear statement of scientific basis and the regulatory purpose and need for the test method. The EDSTAC introduced the Tier 2 tests as having the purpose to “characterize the nature, likelihood, and dose-response relationship of endocrine disruption of estrogen, androgen, and thyroid in humans and wildlife.” Tier 2 tests were designed to be definitive tests which generate sufficient data to characterize the specific hazard of the substance and provide sufficient information on dose-response and adverse effects to

permit risk decisions. **Please comment on the rationale and purpose of the assay as part of the Tier 2 testing in the EDSP, as described in Section 3 of the LAGDA ISR.**

Panel Response

Amphibian development can be used to assay for disruption of gonadal development possibly associated with disruption of androgen or estrogen signaling pathways and for thyroid disruption. The development of validated, reliable methods is critical to the interpretation of individual laboratory findings and understanding where sources of variation may occur.

The Panel agreed that the validation methodology used by the EPA is sound and is based on the ICCVAM and OECD published guidelines. The rationale for the assay was well developed in the LAGDA ISR. It includes the ease of monitoring morphological and growth endpoints in amphibians, the ease of exposure to chemicals, and background knowledge and availability of the sensitive endpoints of gonadal and thyroid histomorphology, plasma VTG concentration and possibly plasma tetraiodothyronine (T4) concentration. Furthermore, the ability to both genetically and phenotypically sex the individual animals is important and increases the utility of the assay. All of these endpoints, although subject to overt toxicity, may also indicate outcomes that involve disruption of estrogen, androgen or thyroid hormone production and/or action. However, although the effects observed may be due to endocrine mechanisms, they may also be due to non-endocrine mechanisms linked to toxic effects on organogenesis. The LAGDA does not allow for determination of the specific MOA for any tested chemical, although some of the endpoints may more accurately predict endocrine-disruptive actions. It is important to recognize that the mechanistic endpoints are biomarkers of exposure, and will not always be diagnostic of a specific MOA. Effects on even seemingly direct measures of endocrine function (thyroid and gonad histology, plasma VTG or plasma T4 concentration may be due to other mechanisms such as disruption of organogenesis that lead to endocrine outcomes.

The Panel recommended that a contingency table of expected outcomes for different modes of action for the LAGDA assay be developed as was provided in the ISR for the medaka assay (see Table 3-4, p. 24, MMT ISR). Such a table would be very useful for distinguishing endocrine from non-endocrine MOAs, or the disruption of specific pathways from toxicity. A description of expected outcomes tied to potential MOAs will increase the utility of the LAGDA.

One of the main goals of the Tier 2 assays is to evaluate endpoints that allow for prediction of population level outcomes. Such outcomes require exposures that are followed through multiple generations, or minimally through reproduction in the first generation that was exposed to the chemical. Unfortunately, multigenerational testing in *Xenopus laevis* is limited by its long generation time (~1.5 years). Another amphibian species that may be amenable to such testing is *X. tropicalis*; however, this species is more difficult to culture in the laboratory than *X.laevis*, and it takes 5-8 months after metamorphosis to reach reproductive maturity. The EPA has identified the problems associated with long-term testing in amphibians and adequately addressed the differences between the two commonly used species of *Xenopus*. The Panel discussed the issue of extrapolating results of the LAGDA to other amphibian species. Chemical effects (or lack thereof) on *Xenopus* species may or may not be reflective

of effects on North American amphibian species. However, the Panel recognized that culture of North American species has not been consistently developed or applied, and therefore, is currently not practical for use in testing protocols. It should be noted though that some species (e.g., the leopard frog) can be bred and reared in the laboratory or cultured in outdoor aquaculture ponds. Problems with understanding negative results will continue for environmental risk assessment given the potential for interspecies variation in sensitivity (see the response to charge question 7 for elaboration). However, the Panel did not recommend expanding the LAGDA to include North American species at this time.

*Question 2. Test methods and their associated endpoint(s) should be scientifically relevant to the biological processes of interest and should be demonstrated to be responsive to the specific type of effect/toxicity of interest. Each species presents unique characteristics from a biological perspective and allows for specialized endpoints to address a specific toxicological mode of action. **Please comment on the biological and toxicological relevance of the assay in regards to the stated purpose of characterizing endocrine disruptors, as described in Sections 3 and 4 of the LAGDA ISR.***

Panel Response

The purpose of the assay is to detect effects of chronic exposure to potential endocrine disrupting chemicals identified in Tier 1 assays on some aspects of amphibian development. Exposure begins during embryogenesis and continues throughout metamorphosis, a period of postembryonic development, and then into the juvenile stage. If the LAGDA can be executed in a reproducible way, it has the potential to have high biological and toxicological relevance for understanding chemical impacts on amphibian development that could also be relevant to other animal taxa. Results from the LAGDA may also suggest endocrine MOA, but it is uncertain whether it will be able to clearly distinguish endocrine from non-endocrine MOA, or from toxicity.

Overall, the biological and toxicological relevance of the LAGDA for detecting developmental abnormalities caused by chemical exposure is strong. Some endpoints may be good measures of endocrine disruption; e.g., thyroid hypertrophy can support a goitrogenic action, elevated plasma VTG can signal an estrogenic action, failure in oviduct formation can signal an androgenic action, and sex reversal may indicate either steroidal or anti-steroid action. But many effects may not be directly related to endocrine disruption per se, but instead due to a failure in organogenesis, either due to the chemical influencing specific cellular pathways, or to toxicity. Therefore, distinguishing a non-endocrine from an endocrine MOA will be difficult, and assessing kidney and liver histopathology will be important to help support a specific (possibly endocrine) MOA vs. toxicity. Concerns were raised by the Panel regarding the small sample sizes and the lack of a power analysis for biologically important differences for endpoints that relate to the specific AOPs, although some Panel members thought that statistically incorporating the individual variation into the models as discussed in the statistics section may help.

Whole organism developmental progression, expressed as the time to reach metamorphic climax (Nieuwkoop-Faber stage 62), is measured in the LAGDA as a proxy for thyroid hormone production and/or action. Concordance among measures of developmental timing, thyroid histopathology and possibly plasma T4 concentration could be a powerful means to support that a chemical has thyroid

active properties. However, results of the validation tests with prochloraz suggest that an effect on thyroid histopathology that was seen in three laboratories (although the LOEC ranged from 6.7 µg/L to 180 µg/L), and that correlated in one laboratory with an increase in T4, was not expressed as a change in development rate. While the LAGDA may be a powerful diagnostic tool to identify chemicals that disrupt thyroid function, findings from the inter-laboratory analysis do not support that the LAGDA has yet been validated for the purpose of identifying chemicals that demonstrate AOPs for animal development via a mode of action that involves the thyroid.

The Panel expected that compounds identified as thyroid disrupters in the Tier 1 amphibian metamorphosis assay (AMA) will also be identified as such in the Tier 2 LAGDA. However, because chemical exposure begins during embryogenesis in the LAGDA, the effects on development may be manifest as a failure of organogenesis independent of thyroid hormone, and any thyroid effects that are seen may be secondary to the teratogenic effects. Therefore, distinguishing an endocrine MOA from a non-endocrine MOA, even if the chemical results in altered thyroid gland development and function, may be difficult. This is also true for the measures of reproductive development.

The LAGDA also monitors gonadal and urogenital ductal morphology and plasma VTG as evidence for disruption of estrogen or androgen production and/or action. Altered reproductive system development is intended to be a proxy for impacts on fertility and fecundity, which are not directly measured in the LAGDA. The oviductal morphology is likely to be the best diagnostic endpoint for predicting effects on reproduction; if there is no oviduct the female cannot oviposit eggs. The Panel indicated that it was not possible to definitively link alterations in testicular morphology or elevated plasma VTG with impacts on fertility or fecundity. Vitellogenin is low or non-detectable in 10 week female frogs. Therefore, while elevation in VTG may be diagnostic of an estrogenic action, this endpoint cannot detect an anti-estrogenic effect. Anti-estrogenic effects may be detected by comparing genotypic and phenotypic sex. Frogs are a good model for this type of outcome since estradiol is necessary for ovarian organogenesis. Male-biased sex ratios can support an anti-estrogenic MOA.

Effects on growth may be due to endocrine or non-endocrine mechanisms; i.e., such effects should not be dismissed as simply representing toxicity, unless they are seen only at the highest concentrations and exhibit clear liver histopathology. Effects on growth could represent endocrine (or non-endocrine) MOA that may or may not have been represented in the Tier 1 assay. The Panel recognized that it may be difficult to distinguish toxicity from endocrine disruption in the LAGDA (or for that matter in all Tier 2 assays). However, this does not compromise the ability to identify adverse outcomes at apical endpoints for chemicals identified in Tier 1 assays as having the potential to disrupt animal endocrine systems. It does compromise the ability to definitively determine the MOA of the chemical.

*Question 3.1. The test protocol should be sufficiently detailed and should include a description of what is measured and how it is measured. The selection of endpoints within the assay should be reflective of the biological processes of interest and the endpoints should be intrinsically relevant and have established sensitivity. The test protocol should demonstrate the ability to measure the endpoints and provide adequate performance criteria for evaluation. **Please comment on the selection,***

optimization and demonstration of the assay endpoints, as outlined in Sections 3 and 4 of the LAGDA ISR.

Panel Response

1) Selection of Endpoints

The Panel indicated that the selected endpoints are appropriate for evaluating chemical effects on growth and development of *Xenopus* larvae, and on gonad and thyroid gland, and urogenital tract organogenesis that may reflect endocrine disruption. Alterations in growth of larvae combined with clear liver and possibly also kidney histopathology may be scored as toxicity. Effects on growth, particularly at lower concentrations of test chemicals, and possibly without clear liver, or kidney, histopathology could indicate disruption of specific cellular signaling pathways (endocrine or non-endocrine).

A change in the timing of metamorphosis combined with thyroid histopathology and possibly changed plasma T4 can signal chemicals that alter thyroid activity. Elevations in plasma VTG in males or females (since females are immature and have low or non-detectable circulating VTG) can signal estrogenic activity as can the presence of testicular oocytes. Abnormal oviduct development can indicate androgenic activity. The oviduct morphology may be the best diagnostic endpoint for potential effects on reproduction. Comparison of phenotypic sex ratios with the genotypic sex ratios is potentially a powerful endpoint for assessing effects of chemicals that may alter sex steroid signaling.

The assay endpoints can support hypotheses for MOA that involve disruption of the endocrine system, but may not be able to distinguish endocrine from non-endocrine MOAs, or the disruption of specific cellular signaling pathways versus toxicity.

2) Optimization

The variability in results among laboratories that participated in the inter-laboratory study suggests that some or all of the methods are not optimized, and/or that the guidelines for their execution can be improved. Assay endpoints appear to be very sensitive to animal husbandry/rearing parameters. The husbandry guidelines for these assays must be established by controlled studies and clearly defined in the protocols. Deviations from the husbandry protocols may disqualify the results.

A source of variation that can impact optimization of the protocol is variation in the rearing water; using local water sources without standardizing for hardness or iodine content (a wide range is allowable under the guidelines) may lead to variation in the response to thyroid hormone disruption and growth parameters. Furthermore, potential endocrine disruptors in source water, if not removed by charcoal filtration, may lead to variation across testing regimes even if other issues like the feeding regime are standardized. Some of the inter-laboratory variation in the LOECs may be related to water quality. For example, mineralization in kidney tubules may be related to water quality.

The Panel noted that there was a range of animal densities used in the different laboratories. This may not be an important issue since as the EPA indicated the flow rate is a more important variable than stocking density. The Panel indicated that the EPA should provide data to support this conclusion.

Impaired growth, combined with liver and kidney histopathology may be robust endpoints for detecting toxicity. However, there are issues related to the histopathology that were identified in the ISR. Gonadal histopathology, particularly testes, appears robust. Ductal histopathology is a particularly important endpoint for supporting that a chemical can impact fertility and/or fecundity, since the potential for population level effects will not be directly addressed in the LAGDA.

There are significant issues with the plasma VTG and T4 assays that are addressed in subparts 2 and 3 of this charge question.

Similar to the other assays evaluated for Tier 2 ecotoxicity screening, the lack of positive controls is a significant limitation to the LAGDA. Positive controls are important for evaluating whether the assay is behaving the way that it should in the contract laboratories. However, there are many potential modes of action represented in the Tier 1 assays and, therefore, selecting an appropriate positive control may be challenging in many cases. In some cases it may be straightforward; e.g., a compound that appears as a goitrogen in Tier 1 can be compared to a well-known goitrogen such as methimazole or propylthiourea in the LAGDA. An apparent estrogenic compound can be compared to estradiol. The positive control chosen will help demonstrate that the assay is working for the predicted MOA, but may not account for MOAs that are not predicted in the Tier 1 assays. Therefore, the positive control may not always provide results in concordance with those found in the test chemical exposure treatments.

The means for choosing the maximum dose of chemicals used in the assays was not clear. Choosing test concentrations that are equally distant on a power or log scale is based on the assumption that the response will be approximately linear when plotted against the test concentration that has been transformed by a power (square root) or to a natural log. With endocrine disrupting chemicals this assumption may not be valid if the endpoint response is U (or inversely U) shaped rather than exponential, or by some power greater than one. It is likely that typical log scale test concentrations will be sub-optimal for identifying U shaped, non-monotonic dose responses.

3) Demonstration

There were four chemicals analyzed in the inter-laboratory validation studies. For some of the endpoints there was consistency among laboratories, at least for the presence of an effect. But the LOECs typically showed great variation among laboratories. For each of the four chemicals analyzed there were significant issues related to reliability and reproducibility of assay endpoints.

Prochloraz is predicted to inhibit aromatase and act as an androgen receptor antagonist. There are recently published data that support that it can impact thyroid function in anuran amphibians (see response to charge question 4). There was large variability among laboratories in the LOEC for thyroid histopathology (the LOECs ranged from 6.7 to 180 µg/L). Effects on gonad and ductal morphology were

more consistent among the laboratories, at least for the males (less so for the females). Plasma VTG could not be used since it was below the limits of detection of the assay.

4-tert-octylphenol is predicted to be a weak estrogen receptor agonist. Effects on thyroid endpoints were seen in only one laboratory (Laboratory A). Laboratory C reported an effect opposite to Laboratory A on time to metamorphic climax, possibly due to problems with animal husbandry. Most laboratories found effects on measures of reproductive development and function, but these were variable among laboratories; some found effects, some not, and the LOECs varied among laboratories when an effect was found. For example, the LOEC varied for effects on duct morphology in females from not significantly different from control to 6.25 to 25µg/L (each laboratory found a different LOEC).

17β-trenbolone is predicted to be a strong androgen receptor agonist. It was tested only in Laboratory A. They found some effects on the testes and male and female ducts at the higher doses tested.

Benzophenone-C is predicted to be an estrogen receptor agonist and inhibitor of thyroid peroxidase. It was tested only in Laboratory E. They reported dramatic effects on thyroid histology at the lowest dose, but increased time to metamorphic climax only at the highest dose (6 mg/L) which exhibited significant toxicity. Similar to the results found with prochloraz exposure, these results question the relationship of thyroid histopathology to thyroid hormone production and the apical endpoint metamorphosis. There were clear effects of the lowest dose of benzophenone-C on reproductive measures, supporting its predicted estrogenic activity.

4) Histopathology

Automated morphometric analysis could allow for quantitation of some, but perhaps not all histological features. It would also provide for greater objectivity in the analysis provided that the choice of histological planes (i.e., their position and the number of representative sections) in relation to the extent of the organ is standardized. For example, the size of organs like the thyroid, testes and the oviduct, and thyroid follicular cell height are potentially amenable to morphometric analysis that could be automated to some degree.

The inter-laboratory analysis showed that there could be divergent results with liver histopathology among the laboratories; this was seen in the prochloraz trial. The basis for these differences may be that two histopathologists conducted the analyses, or that there were animal husbandry issues in the laboratory with divergent results (Laboratory C). The LAGDA ISR authors suggested that this outcome could be due to disease in the Laboratory C animals. The two possible explanations illustrate that a better method or better guidelines are needed for histopathological analyses, and that more comprehensive guidelines are needed for animal husbandry. For the inter-laboratory analysis, it was intended that one histopathologist would analyze tissue sections from each of the participating laboratories. In validating a method that will be applied in different settings, it is important to demonstrate that it performs similarly across laboratories. To assess the validity of the assay, it is important that several histopathologists score the same sets of slides and that the results be compared.

Given uncertainty in the reason for the anomalies, the Panel indicated that it was not possible to determine if the liver histopathology is sufficiently robust and repeatable to be used in the Tier 2 assay.

The assay does not allow for assessment of reproductive behavior, fertility and ultimately fecundity, which is necessary to assess population-level effects.

*Question 3.2. Three of the four labs in the inter-laboratory validation of the LAGDA were unsuccessful in properly performing the thyroxine (T4) analyses using commercially available enzyme-linked Immunosorbent assay (ELISA) kits with antibodies specific for human or canine T4. Subsequently, the EPA developed an extraction method (Section 5.3.2) and presents a revised ELISA method for measuring T4 in amphibian samples in Appendix 7 of the LAGDA ISR. **Please comment on the technical feasibility, reproducibility, and accuracy of the revised ELISA T4 measurement method. Provide any recommendations regarding additional guidance to ensure the reproducibility of T4 measurements.***

Panel Response

The three contract laboratories (ABC Laboratories, Fort Environmental Laboratories, Battelle Marine Science Laboratory) were unable to accurately determine plasma T4 concentrations in metamorphic climax (NF stage 62) stage tadpoles using commercially available enzyme immunoassays (ELISAs) developed for canine or human blood plasma and serum (see T4 study report in Appendix 7, LAGDA ISR). The U.S. Army laboratory study reported plasma T4 concentrations in tadpole and juvenile frogs. However, the measures were highly variable and the values, at least for the juveniles are much higher than plasma T4 concentrations reported in the literature in which radioimmunoassay was used. The Fort Laboratory also measured plasma T4 in tadpoles and juvenile frogs. These values were much lower than those from the U.S. Army laboratory, but were closer to the range reported in the literature for *X. laevis*.

The EPA MED laboratory developed an ethanol extraction protocol to remove interfering substances like plasma proteins that may enhance the ability to measure plasma T4 in tadpoles using ELISA. However, this method is still under development and therefore insufficient information is currently available to determine the technical feasibility, reproducibility and accuracy of the revised extraction and T4 ELISA method.

Members of the Panel questioned the rationale for determining plasma T4 and thus the need for its inclusion in the LAGDA. Beyond the technical difficulties of obtaining sufficient blood from small tadpoles and measuring plasma T4 by ELISA, published findings show that there is often large inter-individual variation in plasma and whole body thyroid hormone levels at NF stage 62 or metamorphic climax. Thus, with the small sample size it may be difficult to detect treatment effects. More importantly, assessing plasma T4 at NF stage 62 may not provide information on impacts on the thyroid function. That the animals have reached NF stage 62 generally indicates that they have produced sufficient thyroid hormone to support tissue transformations. They could have achieved this stage of development through compensatory mechanisms that may be evident in the thyroid histopathology. Hypertrophy of the thyroid may reflect compensation such that sufficient thyroid hormone was

synthesized and secreted to allow the animals to obtain NF stage 62. It is also possible that a chemical will alter tissue sensitivity to thyroid hormone without affecting thyroid hormone production.

Therefore, given the technical challenges in measuring plasma T4 and the limited interpretative power of measurements at the single stage of metamorphosis, NF stage 62, the Panel recommended that the EPA consider eliminating plasma T4 measures from the LAGDA.

Question 3.3. Vitellogenin (VTG) is an established biomarker of estrogenic exposure and is used as a key endpoint in endocrine disruptor testing. There is currently not a standardized commercial source for Xenopus laevis VTG antibodies for an ELISA. 3.3 Please comment on the protocol for measuring and reporting VTG levels. Provide any recommendations regarding additional guidance to ensure the consistency, repeatability and reproducibility of VTG measurements.

Panel Response

Plasma VTG or liver *vtg* mRNA measurements are useful diagnostic tools for disruption of sex steroid signaling, particularly estrogenic signaling. There was large variation in the control plasma VTG concentration measurements among the contract laboratories that conducted the assays. The reasons for the variation are unknown. Most were unable to detect VTG and this is likely due to the immaturity of the frogs at 10 weeks post-metamorphosis. There are two issues to be addressed: 1) the performance of the VTG ELISA and 2) the inability to detect plasma VTG in immature frogs.

Regarding the first issue, the performance of the VTG ELISA, it is difficult to know why the contract laboratories had trouble with the method because assay performance metrics (e.g., data on the standard curves, quality control samples, etc.) were not provided. However, there are some ambiguities in the protocol that could have contributed to the problems:

- 1) The protocol says, “*At the end of the sampling, tubes should be centrifuged to remove the blood cells from the plasma.*” Guidance should be given on the g force generated during centrifugation to avoid hemolysis. It is possible that some of the variance in the assays is due to hemolysis caused by centrifugation.
- 2) The preabsorption of the primary VTG antiserum with control male serum could have introduced a variable that may have led to variation in the assay. It was not clear if this was done at the EPA Duluth laboratory or in the contract laboratories.

Regarding the second issue, the low circulating plasma VTG at 10 weeks post-metamorphosis may allow for sensitive detection of estrogenic compounds compared with mature females. But this needs to be demonstrated.

Other potential sources of variability include:

- 1) Preparation of collection tubes with aprotinin which requires lyophilization in a Speedvac.

- 2) The normal goat serum used for blocking. Different lots may give different results.
- 3) Generating serial dilutions of the top standard as opposed to providing sets of prediluted standards as is typically done in commercial kits.
- 4) Inaccuracies caused by dilution of the samples, particularly in performing the 1:300 dilution that required pipetting only one microliter of plasma.

The ABC Laboratories used a fortified inter-assay quality control sample which was freshly prepared each day of analysis by spiking pooled male *Xenopus* plasma with VTG from the same source as was used to prepare the standards to a nominal concentration of 2 µg/dL. Such QC samples could be incorporated into the VTG ELISA.

However, this same laboratory accepted a CV of up to 20% between assays (the interassay coefficient of variation) which is quite high. Most samples analyzed by this laboratory were below the level of detection (i.e., non-detectable), although they are reported as half the minimum detectable value, presumably to provide a number for statistical analysis. The minimum and maximum detection limits on the different plates varied widely, and samples from a single LAGDA trial were run on different plates on different days. This could introduce an important confound (if the samples are detectable) since there was wide variation in the general parameters of the assay.

Panel members agreed that measuring plasma VTG is an important endpoint for detecting estrogenic activity. This requires a robust assay method that laboratories contracted to conduct the LAGDA can perform. A commercially available *Xenopus* VTG ELISA kit with a standardized protocol for contract LAGDA laboratories to use would be ideal. Alternatively, a contract assay laboratory or laboratories with appropriate expertise could process samples from the contract LAGDA laboratories.

The Panel discussed the possibility of including a measure of liver *vtg* mRNA by reverse transcription quantitative polymerase chain reaction (RTqPCR) in the LAGDA. This maybe a more sensitive biomarker for disruption of estrogenic signaling, but this would need to be evaluated. To the Panel's knowledge, there is currently no validated RTqPCR assay for *Xenopus vtg* mRNA. There are technical challenges that are different in character, but perhaps similar in complexity to the VTG ELISA. For example, RNA isolation, evaluation of RNA quality, reverse transcription and qPCR are technically challenging and may be beyond the expertise of some contract laboratories. A Panel of appropriate reference (e.g., 'housekeeping') genes would need to be identified, assays developed and run in parallel with the *vtg* RTqPCR assay. The concordance between plasma VTG and liver *vtg* mRNA, and the relative sensitivity of both measures to endocrine disruption would need to be evaluated. The SAP recommends that the EPA consider evaluating the feasibility of including *vtg* mRNA in the LADGA.

Question 4. Demonstration of the test method performance should be based on the testing of reference chemicals representative of the types of substances for which the test method will be used. Test substances should adequately represent an appropriate range of responses and physical/chemical properties for which the test method is proposed to be appropriate. The selection of the most appropriate statistical approaches depends in part on the nature of the data and also on the design of the validation study. Statistical and non-statistical methods used to analyze should be described.

Please comment on the selection of test substances and methods (analytical and statistical where appropriate) chosen for the demonstration and validation of the LAGDA assay.

Panel Response

1) Selection of test compounds

Four test compounds were used to validate the LAGDA assay. These are:

- prochloraz, listed in the ISR as an aromatase inhibitor with AR antagonist activity. However, there is at least one previous study with amphibians that shows prochloraz can also disrupt the thyroid endocrine axis and delay metamorphosis (Brande-Lavridsen *et al.*, 2010). The Panel recommended that a thyroid-dependent AOP be added to describe prochloraz in order to fully, and perhaps more correctly, interpret the data collected.
- 4-*t*-octylphenol, listed as a weak ER agonist
- trenbolone, listed as a strong AR agonist
- benzophenone-2 (BP-2), listed as an ER agonist and thyroid peroxidase inhibitor

The Panel did not have concerns with the four compounds chosen. The logic behind the choice of these substances was clear, and each of the compounds is good for proof-of-concept outcomes of the assay. However, because the purpose of these studies was to validate the LADGA assay, additional representative chemicals should be chosen to enhance our understanding of a multitude of AOPs. With this in mind, the Panel questioned the absence of a strong and well characterized thyroid active compound such as propylthiouracil (PTU) or perchlorate for the purpose of unambiguously validating the developmental endpoints of the assay. Also, a strong AR agonist was tested in the assay, trenbolone, but this was not the case for ER agonists, which only included a weak one, 4-*t*-octylphenol. The issue of positive controls to validate the assay is important as has been also pointed out in the discussion of other Tier 2 assays. Finally, the Panel believes the list of test compounds should have included those that were used in the other Tier 2 assays. A greater overlap in test compounds among all the Tier 2 assays would have allowed comparison of assays based on relative performance, and assisted in decisions on which assays would best serve the purpose of a Tier 2 test.

2) Statistical methods

General comments. The Panel had many issues with the statistical methods chosen and described in Section 4.4 of the LAGDA ISR. The Panel believes that these methods were developed some time ago and do not include more recent statistical advances that allow for unequal variance, non-normally distributed data, and modern non-parametric techniques.

In general, the statistical methods should be updated, written more clearly and described in detail for each endpoint and not described as a general set of methods applicable for any endpoint. The type of data collected and experimental unit used for each endpoint were often incompletely identified in the methods, which made it difficult to identify the appropriate statistical approach to use. When combined with the lack of specificity in the statistical methods, this leads to incorrect analyses, ranging from

treating categorical or ordinal data as continuous to failing to include appropriate random or repeated measures effects in the model. For example, StatCHARMS, had code that provided incorrect analyses.

The Panel indicated that the statistical analyses should also include testing across laboratories to identify where differences exist as a means of identifying the more problematic aspects of the testing methods and endpoints.

Specific Comments. The Panel had a range of specific comments on the statistical methods as discussed in Section 4.4 of the LAGDA ISR.

The experimental unit being analyzed was not identified except in a limited number of places, e.g., pooling of plasma for all individuals sampled in a tank in order to analyze T4. In the absence of this information it is assumed that for some of the endpoints it is the tank (e.g., as indicated on p. 19, body weight and snout-to-vent length are tank averages), but it is not clear when individual animals are considered the experimental units for analysis.

A more recent update (LAGDA Errata Validation of the Larval Amphibian Growth and Development Assay Integrated Summary Report.pdf; found in the public e-docket, <http://www.regulations.gov>,) and the answers provided to questions asked during the SAP meeting indicates that the data collected on individual animals are being analyzed as a nested mixed effects model as is appropriate. On the other hand, the presentation provided on Day 1 of the meeting indicated that the mean of the tank was the unit analyzed using a weighted one-way ANOVA with the weights a function of the sample sizes for the means. This later approach is incorrect and should be rerun with the individual values using the mixed effects nested model design. The reason is illustrated by the following model.

$$y_{ijk} = \mu_i + \delta_{ij} + \varepsilon_{ijk}, \quad i = 1, 2, \dots, d; j = 1, 2, \dots, t; k = 1, 2, \dots, n_{ij}$$

where δ_{ij} is a random tank effect and ε_{ijk} is a random error component associated with the animal. So a model for the tank means is given by $\bar{y}_{ij} = \mu_i + \delta_{ij} + \bar{\varepsilon}_{ij}$, $i = 1, 2, \dots, d; j = 1, 2, \dots, t$; In this model, one can show that

$$\text{var}(\bar{y}_{ij}) = \sigma_{\delta}^2 + \frac{\sigma_{\varepsilon}^2}{n_{ij}}.$$

A weighted analysis on “the tank” means, where one weights by the sample size (as is shown in the model above), is not valid unless tank to tank variability measured by σ_{δ}^2 is equal to zero. The Panel indicated that the corrected statistical analysis discussed during the SAP meeting is the approach that should be used.

Assessing normality and heterogeneous variance for such small sample sizes is difficult as the power of the tests is quite low (Razali & Wah, 2011). For example, when the alternative distribution is a slightly skewed distribution, the power is approximately 0.7 for sample size of 50. Transformations or

decisions to use a non-parametric test then become data driven and this potentially impacts the conclusions drawn from the hypothesis testing.

The following statement made in Section 4.4 of the LAGDA ISR (p. 18), “If the data (perhaps after a transformation) are normally distributed with heterogeneous variance, a significant treatment effect is determined from the Tamhane-Dunnett or T3 test or from the Mann-Whitney-Wilcoxon U test”, is incorrect. None of these are tests for an overall treatment effect; i.e., of H_a , at least one treatment mean differs from the other means. Instead, the Panel indicated that a method designed specifically for unequal variances such as Welch’s method (1951) or the Kenward-Roger approach for adjusting the denominator degrees of freedom for the F-test (1997) should be used. Further, the description of tests of differences among treatments for models with normality, but unequal variance, is incorrect. The Tamhane-Dunnett and T3 tests are correctly identified as methods for controlling experiment-wise error rate when performing multiple pairwise t-tests of differences between two means with unequal variances; whereas, the Mann-Whitney-Wilcoxon U test is not correct. The Mann-Whitney-Wilcoxon U test is a non-parametric version of a t-test, not a method for controlling experiment-wise error rates. In addition, using non-parametric tests, such as the Kruskal-Wallis test for a non-parametric one-way ANOVA, with these small sample sizes also has low power.

The Panel indicated that the EPA’s recommendation to perform pairwise testing of means when the omnibus test (i.e., the F-test or non-parametric version) fails to reject the null hypothesis must be carefully considered and described. If the experiment-wise error rate is controlled for the multiple comparisons, such as is done with Dunnett’s test against a control or the Tukey-Kramer (Tukey, 1953; Kramer, 1956) approach for pairwise comparisons, then ignoring the overall F-test is not an issue, but if the method for performing the pairwise tests does not control experiment-wise error rates, such as the Mann-Whitney-Wilcoxon U test without adjustment, then the type I error rate is larger than the nominal rate given and the conclusions will be erroneous.

The Panel stated that the use of the median from the Kaplan Meier (KM) product estimator as the response variable in an ANOVA is inappropriate because the variance associated with the estimated median is ignored in the subsequent analyses. The Panel recommended that an alternative be used, such as the log-rank test that compares the KM survival curves directly with appropriate adjustments for multiple comparisons of the curves, such as the Sidak correction test discussed in the U.S. Army Laboratory LAGDA validation study (study found in the regulatory e-docket located at <http://www.regulations.gov>, docket: EPA-HQ-OPP-2013-0182, document number: EPA-HQ-OPP-2013-0182-0041.pdf).

The Panel also disagreed with the EPA’s recommendation to use a statistical test based on the observed results such as what is described in the LAGDA ISR (p. 19), “*a step-down Jonckheere-Terpstra trend test should also be applied to the data as long as a monotonic response was observed.*” This test inflates the type I error rate above the nominal and should have been decided *a priori* not *a posteriori*.

The Panel noted that mortality, expressed as the number dead among the number observed for some given time period, should be analyzed as a binomial random variable with a generalized linear model

(one-way ANOVA with a binomial response), not as approximately normal after transformation. The arcsine square root transformation ignores the number of individuals that were observed and treats what is a categorical variable as continuous. Further, it is best suited to data with large number of observations (more than 10 juveniles) and percentages near 0.5. It is not a good transformation when the percentages are near 0 or 100. An alternative, if one insists on transformation of the data, would be to use the logit transformation (i.e., the log of the ratio of p to $(1-p)$).

The use of the RSCABS method for severity scores used in the histopathology analyses might be appropriate (see previous discussion in “General Comments Concerning Statistical Issues”). One concern with this approach is that it assumes that any effect of concentration must be monotonic.

From the descriptions provided by each laboratory performing the tests, the Panel concluded that the protocols were not always followed to the letter. For example, the number of larvae per tank ranged from 20 to 30 depending on the laboratory, the test substance concentrations were 20 $\mu\text{g/L}$ or 25 $\mu\text{g/L}$ for 4-*t*-octylphenol and 20 $\mu\text{g/L}$ or 30 $\mu\text{g/L}$ for prochloraz, and the volume of water/tank varied from 4-L to 10-L. In addition, the nominal and actual concentrations varied among laboratories as well. Most of the laboratories failed to meet the condition of low variability ($\leq 20\%$ CV) and the actual concentration is within some reasonable value to the nominal.

One important reason for the lack of consistency in results among the laboratories is the sample sizes are insufficient to observe effects, if they do exist, because of the lack of power. The Panel strongly recommended that the results obtained to date be used to do a power analysis for identifying a better experimental design. For example, there should be an effort to identify the contributions of the various sources of variance (within tanks versus among tanks, the effect of inconsistent treatment concentrations over time, the effect of mortality on the variance estimates, etc.). This information should be used to identify the optimal number of larvae or juveniles per tank and the number of tanks. Of course, specific endpoints may require different sample sizes; therefore, a study should also be done to identify the most important endpoints and necessary sample sizes to see differences among concentrations, if they exist. All of these analyses for sample size should also take into account the size of biologically relevant differences and whether the treatment concentrations of the tests substances were the best set of comparisons. The Panel suggested that consideration be given to whether covariates should be added into the modeling to account for some of the sources of variation that are observed in the inter-laboratory studies. However, the low sample sizes probably preclude this approach.

Overall, the Panel concluded that the statistical methodology is not readily transferable because there appears to be several instances where the protocol was not implemented as planned. Further, there is a serious issue with the failure to identify laboratory methods for some of the endpoints (e.g., the ELISA for T4 and VTG). In addition, the lack of consistency among pathologists argues for a validation exercise or training mechanism that can provide reproducibility among pathology readings by different scientists. See discussion on transferability in the response to charge question 6.

Question 5. Considering the variability inherent in biological and chemical test methods, a test method needs to be repeatable and reproducible. A test is robust and reliable if the results are

repeatable and reproducible within a laboratory and between different laboratories, respectively. A test protocol should provide sufficient guidance to ensure proper and consistent performance across labs and chemicals. Please comment on the test method robustness and reliability and the repeatability and reproducibility of the results obtained with the LAGDA assay.

Panel Response

For the purposes of the response to this charge question, the Panel is using the same definitions of reliability, repeatability and reproducibility described in the response to MMT charge question 5. Each of the reference chemicals used for assay validation was associated with one or more endocrine-dependent AOPs discussed in the ISR. At a basic level, the general reliability of the assay can be initially evaluated by comparing effects of each test chemical against expected endpoint results of their respective AOPs.

With the exception of prochloraz, which had some unexpected effects on the thyroid in relation to the AOPs described in the LAGDA ISR, the LAGDA assay generally, though inconsistently with respect to dose outcomes, performed as anticipated for the reference materials. Prochloraz was expected to exhibit aromatase inhibitor and AR antagonist activity. Among the reproductive endpoints, this compound caused changes in overall gonadal and duct pathologies in juvenile males and females consistent with these AOPs. The unexpected result was the most consistent finding (by 3 of 4 laboratories) of thyroid axis disruption effects, including thyrocyte hypertrophy and hyperplasia. As was noted earlier, there is at least one previous study with amphibians showing that prochloraz can disrupt the thyroid endocrine axis and delay metamorphosis (Brande-Lavridsen *et al.*, 2010). The Panel recommended that the Agency add a thyroid-dependent AOP to this compound to include the new information and eliminate the qualification of the thyroid results as “unexpected.”

The Panel also evaluated the endpoint outcomes for the other tested compounds. Although the weak ER agonist used in the assay, 4-*t*-octylphenol, did not yield consistent observations with plasma VTG, when effects were reported, they were positive. Similarly, effects on duct development, especially in females, were also noted that are consistent with the ER agonist activity of this compound. Trenbolone, a strong AR agonist, resulted in reproductive effects (duct development) in both males and females that were consistent with its AR agonist activity. Finally, benzophenone-2 (BP-2), which exhibits ER agonist and TPO inhibitory activities, strongly increased plasma VTG in both males and females and had gonadal and duct pathologies consistent with its ER agonist activity, including 100% sex-reversal of genotypic males at the highest concentration. This compound also delayed metamorphosis and affected thyroid endpoints in a manner consistent with its thyroid peroxidase inhibitory activity. Overall, the general concordance between expected and actual outcomes of the LAGDA assay, when tested with compounds that utilize different MOAs, provides a basic level of confidence in its reliability.

To more rigorously address the question of reproducibility, the ISR presented the results of two sets of ring tests, one with prochloraz involving four different laboratories and another with 4-*t*-octylphenol involving three laboratories. In the ring test using prochloraz, the only individual endpoint that showed a consistent positive response across the four laboratories was a concentration-dependent delay in the

involution of the oviduct in juvenile males. It is also noteworthy that the LOECs for this endpoint were comparable across laboratories. Among consistent “negative” results with prochloraz, all laboratories recorded no delays in the time to NF stage 62, no effects of oviduct development in females, and in all cases the genotypic sex matched the gonadal phenotype of individuals, thus indicating the absence of sex reversal. However, technical problems associated with histological sample processing/reading and the analysis of T4 prevented a full assessment of inter-laboratory variability for many of the apical endpoints associated with endocrine pathways.

In the second round of inter-laboratory assays using 4-*t*-octylphenol, observations of thyroid histopathology or metamorphosis yielded inconsistent results among laboratories; i.e., comparisons of T4 could not be conducted because of technical problems in some of the laboratories. Also, although female duct development was consistently accelerated in a concentration-dependent fashion by 4-*t*-octylphenol treatment across all laboratories, there was a 10-fold difference between the smallest and the largest LOEC reported. Among the consistent negative results, genetic sex matched gonadal phenotype, showing no sex-reversal.

Because the animals are not raised to reproductive maturity and tested for endpoints associated with fecundity and fertility, data will suggest, but not definitively answer, any outcomes with regards to long term adverse effects. Specifically, findings of gonadal or thyroid-related disruption through the juvenile stage will not be able to predict whether there are long term adverse health outcomes on adults following developmental exposure.

Overall, the inter-laboratory testing yielded mixed results. Very few endpoints showed consistent results across laboratories, especially for the thyroid and developmental endpoints and, among the reproductive endpoints, the oviduct was the only endpoint in males and females where a level of consistency was demonstrated. The ISR noted that some of the inter-laboratory variability may have been due to technical difficulties with the procedures experienced in some of the laboratories or their failure to abide by test guidelines. This may indicate a problem with the transferability of the assay in its present form. The Panel concluded that reliability and robustness have not been fully demonstrated for the LAGDA assay.

*Question 6. The test protocol should be descriptive enough to be fully transferable to a functional laboratory. The protocol should describe the methodology of the assay in a clear and concise manner so that a laboratory could comprehend the objective, conduct the assay, observe and measure prescribed endpoints, compile and prepare data for statistical analyses, and report the results. Sections 5 and 6 of the LAGDA ISR outline the process of and challenges experienced in the inter-laboratory studies. **Please comment on the transferability across labs and provide any suggestions or recommendations for improvement of the LAGDA assay.***

Panel Response

There were significant issues with transferability of techniques across laboratories. Only rarely did results from some endpoints corroborate among the laboratories that participated in the inter-laboratory

analysis. Also, effects on liver and kidney histopathology, which was the most consistent finding across laboratories, often exhibited large differences in the LOECs. When results were similar across two to three laboratories, the LOEC values differed by more than 10-fold. For example, in the cross laboratory comparison following prochloraz exposure, three of the six reproductive endpoints exhibited effect outcomes consistently across all four laboratories, but the range of LOECs varied by 20-fold. Only one endpoint, the delay in oviduct involution in juvenile males, was fairly consistent across the laboratories. Because differences in outcomes were more common than similarities, the ability to transfer the technology across laboratories is questionable.

The Panel made the following recommendations to enhance transferability of LAGDA:

- 1) The protocols need to provide step-by-step details on husbandry, water quality, treatment exposure, tissue collection, quantification of morphological endpoints, histopathology, biochemical assays and statistical analysis.
- 2) Because the laboratories conducting the LAGDA procedures at times failed to follow the specific protocols provided by the EPA, the protocols should be more detailed and specific. In addition, directed training should be provided to any laboratory contracted to run the LAGDA.
- 3) As stated for the other Tier 2 ecotoxicity tests, positive controls should be included in the LAGDA. The positive controls should be run in all laboratories that participate in inter-laboratory validation studies and routinely incorporated by any contract laboratories hired to run the assay. The controls may be run at only one dose determined to be consistently different from controls in the validation process. Some specific suggestions include addition of T4 for thyroid endpoints, estradiol or ethynyl estradiol for estrogenic endpoints and trenbolone for androgenic endpoints. These are strong agonists for the pathways being tested through the EDSP and should provide fairly consistent results across laboratories (i.e., T4: increased rate of metamorphosis and thyroid histopathology; estradiol: sex reversal and/or gonadal histopathology; trenbolone: urogenital duct pathologies). Additionally compounds that act as antagonists to these hormonal systems may also be useful. Having at least one of these compounds used in each run of the LAGDA (chosen based on the potential MOA detected in the Tier 1 assays) will provide quality control to insure the assay is functional in the hands of the contract laboratory.

Question 7. The purpose of the validation process is to determine the readiness of a test for inclusion in a testing program. A component of readiness of a test is the evaluation of the usefulness and limitations of the test, including the classes and types of test substances that can and cannot be tested. Please comment on the strengths and/or limitations of the LAGDA assay.

Panel Response

Xenopus laevis is a well-established vertebrate model system for developmental biology. This species has also been used as a model organism for studies of endocrine disruption because of the ability to easily evaluate endpoints of gonadal, urogenital tract and thyroid hormone-related developmental processes. These endpoints are a particular strength of this assay. Although it will be

difficult, if not impossible, to determine the precise MOA of a test chemical using the LAGDA assay, the results may support an endocrine MOA identified in the Tier 1 screen. The potential for disruption of growth, development and/or reproduction are important endpoints for evaluating AOPs and the LAGDA assay, through evaluation of gonadal and urogenital duct development and the combined endpoints of time to reach specific metamorphic stages and thyroid histopathology, will allow for the possibility of inferring AOPs.

The LAGDA ends at the juvenile stage. Thus, while it may be possible to infer AOPs based on this assay, it is not possible to conclude whether a chemical can impact fertility or fecundity and whether it has the potential for population-level effects. The one exception is when urogenital duct development is completely inhibited. Thus, the LAGDA may suffer from false positives (i.e., apparent effects on gonadal, urogenital tract or thyroid development that do not translate into altered fertility or fecundity) and also from false negatives (i.e., no overt histological changes, but significant impacts on reproduction).

The EPA charge question document specifically states that the purpose of the Tier 2 assays is to “characterize the nature, likelihood, and dose-response relationship of endocrine disruption of estrogen, androgen, and thyroid in humans and wildlife.” While the Panel supports the EPA’s logical development and validation of this assay to assess the likelihood of endocrine disruption, the inter-laboratory validation studies provided to the Panel do not demonstrate a reliable and repeatable dose-response relationship and LOECs. Nevertheless, a few endpoints that are related to AOP’s were remarkably consistent, and provide especially useful information in a Tier 2 approach (such as ductal development outcomes). For many other endpoints there was significant variation among laboratories with some finding effects and others finding none. Variable results across laboratories suggest that there will be considerable risk of false negatives or positives in implementing LAGDA. Given the inter-laboratory study variation in LOECs found when using *X. laevis* as a model, suggests that extrapolating from *X. laevis* to other wildlife species will have high uncertainty. A number of studies have demonstrated there is high variation in the toxic susceptibility to exposure (Junges et al., 2012; Distel & Boone, 2011; Garcia-Munoz et al., 2010). In the absence of reasonable certainty, a margin of error for exposure levels should be considered as a buffer against the potential for adverse effects on natural populations.

In summary, the LAGDA, as a tool for risk assessment of endocrine disrupting chemicals, has the potential to identify the risk of different adverse outcome pathways in *X. laevis* that may be applicable to other wildlife. Unfortunately, high inter-laboratory variation indicates that the validation efforts have not demonstrated the ability to achieve consistent results across laboratories. Without reliable results in the LAGDA making broader extrapolation of the results to wildlife species is difficult. The Panel had a range of specific comments on and recommendations for the statistical methods (see detailed response to charge question 4). The Panel recommended that clearer and more stringent protocols and QA/QC parameters for performance should be developed and implemented, positive controls should be incorporated into each run of LAGDA and more appropriate application of statistics including power analysis should be implemented into the LAGDA protocol before this test may satisfy the stated purpose as a Tier 2 ecotoxicity test. A detailed discussion of all statistical issues and recommendations for all four Tier 2 assays is found in the section entitled, “General Comments Concerning Statistical Issues”.

Mysid Two-generation Toxicity Test (MTTT) and Harpacticoid Copepod Development & Reproduction Test (HCDRT) Charge Questions

Question 1. A rationale for the test method should be available, including a clear statement of scientific basis and the regulatory purpose and need for the test method. The EDSTAC introduced the Tier 2 tests as having the purpose to “characterize the nature, likelihood, and dose-response relationship of endocrine disruption of estrogen, androgen, and thyroid in humans and wildlife.” Although the hormones produced and used by invertebrates are not directly analogous to those of vertebrates (e.g., estrogen, androgen, and thyroid), growth, reproduction, development, and other aspects of invertebrate physiology and life cycle are known to be under endocrine control. EDSTAC went on to note that “chemicals that affect these vertebrate hormones may also affect invertebrate hormones resulting in altered reproduction, development, and growth.” Tier 2 tests were designed to be definitive tests which generate sufficient data to characterize the specific hazard of the substance and provide sufficient information on dose-response and adverse effects to permit risk decisions. Please comment on the rationale and purpose of the assay as part of the Tier 2 testing in the EDSP, as described in Sections 2.1 and 2.2 of the MTTT ISR.

Panel Response

The Panel endorsed the concept of the mysid two-generation test as a Tier 2 assay and indicated that this assay is a very sensitive test to measure the possible environmental impacts of toxic or endocrine disruptive compounds on estuarine invertebrates. Invertebrates are significant parts of global ecosystems and thus must be protected from adverse effects of chemicals that disrupt their endocrine systems. Invertebrate endocrine systems will respond quite differently than vertebrate endocrine systems; thus, they need to be addressed in terms of the ecotoxicological risk assessment for endocrine disrupting chemicals. Both the mysid and harpacticoid copepod assays are relevant assays for addressing endocrine disrupting chemical effects in invertebrates.

The Panel noted that the initial EDSTAC charge for developing Tier 2 Tests was focused on “characterizing the nature, likelihood, and dose-response relationship of endocrine disruption of estrogen (E), androgen (A), and thyroid (T) pathways in vertebrates including humans and wildlife.” However, the EDSTAC clearly defined the linkage between endocrine disruption in vertebrates and invertebrates by stating that “further assessing the effects of chemicals that disrupt E, A, and T in vertebrates should be further investigated in invertebrates for disruption of growth, development and reproduction.” The current mysid and copepod bioassays described in the MTTT ISR are designed specifically to address alterations in survival, growth, development and reproduction to indirectly define endocrine disrupting chemical effects in invertebrates. The Panel indicated that what is missing in the mysid and copepod bioassay approaches is a clearly defined endocrine disruption MOA test endpoint, such as disrupting ecdysone or other important invertebrate hormones, which would be analogous to the E, A, and T endpoints in vertebrate testing. Further, the Tier 2 invertebrate bioassay makes no attempt to quantify the molecular/biochemical effect relying, instead, on individual level responses. This makes it

impossible to embrace an AOP approach for invertebrates that is parallel to that done for vertebrates, making conclusions drawn from these bioassays equivocal at best. If an invertebrate AOP approach is developed to complement the current E, A, and T AOP approaches, it will be important to consider a large variety of invertebrate endocrine disrupting chemical pathways that should be protected, not only in crustaceans, but also in mollusk and other important classes of invertebrates. This is particularly important given the ecological importance of invertebrates and their commercial and recreational importance. As a result, multiple invertebrate species should be considered as EPA has done.

The MTTT ISR has a clear description of why invertebrates should be considered in endocrine disruptor chemical screening due to the significant ecological linkages between invertebrates and vertebrates in terrestrial, aquatic, estuarine and marine ecosystems. The MTTT ISR indicated the importance of the linkage of the mysid test to estuarine ecosystems environmental protection. For example, in estuarine ecosystems most of the estuarine-dependent, commercially and recreationally important finfish species (86%) begin their earliest life history stage in the estuary, initially consuming copepods followed by amphipods, mysid shrimp, grass shrimp and penaeid shrimp as they mature. This “crustacean cascade” is a vital link to the functioning of estuaries as nursery grounds for vertebrate, finfish species. Since 53% of the US population lives in the 17% of the land area adjacent to the coast, there are 5.19 times more people per square mile and 5.87 times the gross domestic product per square mile. This situation results in significantly more pollutant burden in these coastal ecosystems. As a result, organisms in coastal ecosystems have significant potential to be exposed to contaminants including endocrine disrupting chemicals.

The current battery of invertebrate Tier 2 tests use crustaceans as model invertebrate organisms that were chosen because of their ecological importance and sensitivity to contaminants. Crustaceans are generally more sensitive than other invertebrates in terms of conventional acute and chronic toxicity testing (e.g., Harmonization of FIFRA and Clean Water Act (CWA) Testing Protocols). Crustaceans have played a critical and effective role in characterizing the potential effects of known invertebrate endocrine disrupting chemicals resulting in effective regulation of insect growth regulators (IGRs) including methoprene, dimilin and fenoxycarb. The conventional, one generation mysid test has played a key role in the risk-based approach for regulating these endocrine disrupting chemicals by establishing a clear dose-response relationship as prescribed by the EDSTAC process. While the mysid test provides a clear assessment of ecological effects to aquatic organisms, it is less indicative of benthic dwelling organisms. Since many chemical contaminants in aquatic and estuarine/marine ecosystems partition into sediments, the Panel suggested that a benthic organism be added to these tests. The addition of the harpacticoid copepod (*Amphiascus tenuiremis*) test provides an additional assay for assessing the impacts of endocrine disrupting chemicals in invertebrates and addresses the concern of sediment. Further, the copepod assay was initially developed as a sediment toxicity test and has been shown by EPA (Fulton et al., 2006) to be one of the most sensitive toxicity test species to a variety of chemical contaminants (e.g., pesticide, PAHs and trace metals).

The MTTT ISR states that “Tier 2 tests were designed to be definitive tests which generate sufficient data to characterize the specific hazard of the substance and provide sufficient information on dose-response and adverse effects to permit risk decisions.” Both the copepod and mysid bioassays address

these tenets of Tier 2 test requirements, but are compromised by the lack of identification of direct endocrine disrupting chemical effects. These tests include growth, survival, development and reproduction endpoints, which make it more difficult to be certain that any of these effects are linked to endocrine disruption as opposed to overt toxicity or due to energetics related stress. Concern over this limitation was presented by the Endocrine Policy Forum during the meeting (comments are available in the public e-docket located at: <http://www.regulations.gov>, docket: EPA-HQ-OPP-2013-0182). The Panel stated that consistency in the results across laboratories was low. This evidence suggests that a given result or set of results may not be repeatable. The lack of consistency across all laboratories and between F1 and F2 generations is also indicative of difficulties in conducting these assays and defining the dose-response effectively. The Panel was very concerned that not all participating laboratories were able to meet control performance criteria and that inter-laboratory variability was significant in both assays. See the responses to charge questions 5 and 7.

The Panel supported the use of a Tier 2 invertebrate assay, but was concerned by the lack of a Tier 1 screen for invertebrates in the current EDSP and expressed concern over using identification of E, A, and/or T active compounds in vertebrate screens to trigger a Tier 2 invertebrate screen. Without a Tier 1 invertebrate assay that focuses on a chemical's potential to interact with the invertebrate hormonal pathway (e.g., ecdysteroids), the Tier 2 test is out of place and its role in the overall endocrine screening program uncertain.

The Panel recommended the EPA consider whether any of the current invertebrate Tier 2 screens could be effectively modified to develop a Tier 1 screen. For example, the addition of a molecular endpoint (invertebrate hormone) to the conventional one generation mysid test might be a useful Tier 1 screen. Alternatively, the copepod bioassay would be a useful screening assay given its miniaturization and comparative sensitivity with the mysid bioassay recommended for Tier 2 by EPA.

Conclusions and Recommendations

The Panel endorsed the concept of the mysid two-generation test as a Tier 2 assay and indicated that this assay is a very sensitive test to measure the possible environmental impacts of toxic or endocrine disruptive compounds on estuarine invertebrates. The data in the MTTT ISR showed that the extended time required for producing the F1 generation broods was more protective (usually due to growth, sometimes reproduction) 30% of the time, 70% of the time the F0 endpoints were equally or more sensitive to F1. Only once did the F1 generation reproduction (brood size) serve as the source of significant difference from the controls (ketoconazole). Therefore, given the small number of compounds tested for the validation testing, the Panel suggested that a great number of more sensitive endpoints be identified in the F1 and F2 generations that would provide greater scientific justification for conducting additional assays. The proposed Tier 2 copepod test is often more sensitive than the mysid assay and some distinct advantages over the mysid assay. Thus, the Panel concluded that both assays have an important role to play in EDSP Tier 2 testing.

Although the Panel was appreciative of the importance of including invertebrates in the EDSP Tier 2 ecotoxicity tests, and acknowledged that some compounds (e.g., general anti-steroidogenic compounds,

such as prochloraz) could have broad effects on both invertebrates and vertebrates, the use of the Tier 2 mysid two-generation test without a representative Tier 1 test or Tier 2 pre-screen results will provide less than objective conclusions regarding a chemical's endocrine disrupting potential and is counter-intuitive to the current EDSP framework. This is due both to the lack of guidance on the approach for deciding which chemicals following the Tier 1 screening battery of tests will undergo a Tier 2 assay and due to the difference in the actual hormones and endocrine receptor control of reproduction and development in invertebrates compared to vertebrates. Therefore, given the universal importance of invertebrates in healthy ecosystems of all kinds, the Panel recommended the EPA include ecdysteroid controlled processes as part of the EDSP Tier 2 assays, along with estrogen/testosterone/thyroids to help build the consistency in the framework between the Tier 1 and Tier 2 testing scheme.

*Question 2. Test methods and their associated endpoint(s) should be scientifically relevant to the biological processes of interest and should be demonstrated to be responsive to the specific type of effect/toxicity of interest. Each species presents unique characteristics from a biological perspective and allows for specialized endpoints to address a specific toxicological mode of action. **Please comment on the biological and toxicological relevance of the assay in regards to the stated purpose of characterizing endocrine disruptors, as described in Section 2 of the MTTT ISR.***

Panel Response

In general, the Panel stated that both the mysid and copepod assays have high biological and toxicological relevance for helping define endocrine disrupting chemical relevancy. Both assays have multiple and inter-related survival, growth, development and reproductive endpoints that are useful in describing whether or not a chemical identified in the Tier 1 screening battery as altering E, A or T AOPs will also affect invertebrate-specific AOPs. While E, A and T are lacking in invertebrate endocrine systems, there are several possible outcomes that may occur in Tier 2 screening of invertebrates including: 1) the compound may adversely affect some invertebrate hormone mediated pathway, 2) the compound may cause overt toxicity through a non-invertebrate endocrine mediated mode of action, 3) the compound may affect both an invertebrate hormone mediated pathway and cause overt toxicity through some non-invertebrate or endocrine disrupting chemical mediated pathway or 4) the compound may not illicit any adverse effects in invertebrates. The conclusion of one of these outcomes over the others may be difficult and have high levels of uncertainty.

The assay is appropriately designed to determine endocrine mediated processes such as reproduction and development as well as the duration of time to reproduction (a measure of growth and maturation rate). Weight and length are good measures, but the Panel questioned whether they have to be measured daily. The Panel noted that there was no measure of time between molt stages, which would indicate impact on one of the most understood endocrine-related processes in invertebrates (e.g., ecdysis). Weight and length at day 7 and 14 are good measures of the summative effects of multiple molt stages that would indicate direct impacts on ecdysis. However, apical measures used alone without a Tier 1 test would not identify a compound as an endocrine disrupting chemical without direct MOA data.

The Panel recommended that the EPA consider adoption of a mysid ecdysteroid receptor assay such as that developed by Yokota *et al.* (2011) or Gaertner *et al.* (2012). Currently, these assays do provide data to predict population-level effects based upon the alterations in growth, development and reproduction. These results have been and are very useful for protecting insect and possibly other invertebrate populations. The population modeling ability of the copepod assay is directed to take test endpoints to predict multigenerational population responses and the mysid multigenerational assay can be used in a similar manner.

The Panel and one of the public commenters indicated that current endpoints are also provided by the current mysid chronic one generational study that is routinely conducted for most chemicals (see public comments by Dr. John Brausch on behalf of the Endocrine Policy Forum, found in the public e-docket, <http://www.regulations.gov>, docket: EPA-HQ-OPP-2013-0182). Results for chemicals run in the proposed Tier 2 tests, where there is also data for the same chemical tested using the conventional mysid one generational assay, should be compared using population models for concordance. Based upon this approach, the Panel indicated that the proposed Tier 2 invertebrate tests have relevance as they are meaningful and useful for the particular purpose of defining impacts on population level, growth, development and reproduction endpoints that may be related to endocrine disrupting chemical effects.

The mysid shrimp bioassay has been used in regulatory toxicity testing for pesticides and other chemicals for more than 30 years. The current mysid shrimp standard testing protocols are designed to assess a variety of toxicity, growth, development and reproductive endpoints that are integrated to evaluate the neuroendocrine control of the tested chemical (e.g., ecdysone). One shortcoming of this approach for assessing the chronic effects of endocrine disrupting chemicals is that these combinations of growth, development and reproductive endpoints may be affected by both endocrine and non-endocrine chemical effects. What is missing in the current mysid test protocol is a direct measure of an endocrine disrupting chemical effect, such as ecdysone or other juvenile hormones controlled by the invertebrate endocrine system, which would be analogous to E, A and T in vertebrate systems.

At present, there are no ideal model biomarkers of invertebrate endocrine disrupting effects; therefore, the survival, growth, development, and reproductive endpoints in the current mysid test are used. Current research is underway to measure ecdysone and other aspects of juvenile hormonal control of molting, development, growth and reproduction. Results from this research indicate that rapid tests for an endocrine disruption specific MOA in invertebrates could be developed and implemented, at least with respect to ecdysone-related mechanisms. See the list of assays in the “Other comments” section of this response.

Despite these limitations, the conventional mysid test has been used to safely and effectively assess the risk of known invertebrate endocrine disrupting chemicals registered for use as pesticides including ISRs, such as methoprene, dimilin and fenoxycarb. Given the success in using the mysid assay in regulating IGRs, the Panel recommended the EDSP include a wider variety of chemicals which may disrupt invertebrate growth, development and reproduction.

Following are strengths and weaknesses identified by the Panel for each of the current Tier 2 invertebrate tests:

1) Mysid Test Strengths and Limitations are fully described in response to charge question 7.

2) The Harpacticoid Copepod Test. The Harpacticoid Copepod Test is an additional bioassay that been developed to assess endocrine disrupting chemical multigenerational effects in invertebrates. This assay and the Mysid Multigenerational Test have been recommended by OECD for inclusion as Phase 5 tests in the OECD invertebrate endocrine disrupting chemical toolbox. Some of the major strengths and weaknesses of this copepod assay are listed as follows:

Copepod Test Strengths

- 1) Test has the ability to assess survival, growth, development and reproductive effects both at an individual and population level for most end points. This may enhance the ability of the assay to discern invertebrate endocrine disrupting chemical effects (as was stated by EPA in their response to questions following their presentation)
- 2) Copepod test can be used to assess aquatic risk, but also has the potential to assess sediment toxicity. However, sediment toxicity can only be done in a flow thru system with beakers which pool males and females and would lose the individuality of the current aquatic, static test.
- 3) Statistical approaches are well described and allow for development of population modeling based on alterations in test end points.
- 4) Test can be run in 96 well plates and can be used in smaller plates.
- 5) Test generates about 1 gallon of waste/test. If 1000 compounds are tested then there will be 1000 gallons of waste. Each drum contains 55 gallons. $18.2 \text{ drums of waste} \times \$1000/\text{drum} = \$18,200$ total costs.
- 6) Copepod test is recommended as part of the OECD testing protocol (Level 5).

Copepod Test Weaknesses

- 1) Copepod test is not used as widely as the mysid bioassay and is not currently used for environmental compliance *per se*.
- 2) Copepod test is not conducted by current pesticide testing laboratories so training will be needed.
- 3) Test delivery of target dose is more difficult given the smaller bioassay test design, which reduces the ability to measure actual exposure concentrations.
- 4) There are some problems in distinguishing between some life history stages.

Conclusions

The Panel stated that both the mysid and copepod Tier 2 assays should be used to quantify interactions with the endocrine system and provide data to predict population-level effects. As such,

the current multigenerational mysid and copepod assays fall short in directly quantifying interactions with the endocrine system because there is little known about the endocrine system of mysids, in particular, and invertebrates, in general. Growth and development, which are some of the key test endpoints, are tied to molting, which is controlled by invertebrate hormones. Both protocols do not directly measure molting or the hormone(s) that control this activity. The Panel endorsed the concept of the mysid two-generation test as a Tier 2 assay and indicated that this assay was a very sensitive test to measure the possible environmental impacts of toxic or endocrine disruptive compounds on estuarine invertebrates. The harpacticoid copepod test is an additional bioassay that has been developed to assess endocrine disrupting chemical multigenerational effects in invertebrates. The Panel recommended the use of the OECD toolbox approach, which includes both the mysid and copepod multigenerational tests that can be used to protect the large diversity of invertebrates.

Other Recommendations

The Panel made the following recommendations to improve the tests:

- 1) Addition of molecular pathway indicators and metabolomics endpoints.** Addition of molecular and metabolomics endpoints may help better define the linkage between growth, development, survival and reproduction with endocrine disrupting chemical related mechanisms of action. For example, the National Oceanographic and Atmospheric Administration (NOAA) has generated microarrays for oysters and crustaceans and has developed metabolomics endpoints in dolphins and crustaceans that would help better define the underlying gene expression. The Panel encouraged this type of molecular based focused research, which may be useful in better defining linkages to invertebrate endocrine effects. Defined linkages could be used to add relevant endpoints in the Tier 2 assays, and for possible Tier 1 invertebrate assay development. There have been studies that have addressed the isolation and genetic expression of the invertebrate hormones as rapid screening tools. Gaertner et al. (2012) reported that copepods exposed to fipronil, and potentially other endocrine disrupting compounds, resulted in impacted reproduction that was linked to a significant increase in ecdysone receptor transcriptional expression at 30 hours compared to controls. Ecdysone receptor transcriptional measurement has been demonstrated as a sensitive and rapid biomarker of ecological relevance when linked to traditional *A. tenuiremis* bioassays (Gaertner et al., 2012). Additionally, ecdysone receptor binding assays have recently been developed for mysids which would allow rapid screening of suspected endocrine disrupting chemical compounds for ecdysteroid specific MOA (Yokota *et al.*, 2011; De Wilde et al, 2013).
- 2) Use of positive controls.** Future inter-laboratory calibration exercises should include positive controls. Positive controls will add cost but are needed to assure that contract laboratories are adequately performing the assays. Positive controls should be added using methoprene or another known invertebrate endocrine disrupting chemical, which would be initially run at a single dose with each new compound tested. Some thought should be given to the idea of positive controls which may target different aspects of invertebrate endocrine pathways. As more and more data are generated for specific invertebrate endocrine disrupting chemical pathways represented by positive controls, the QA/QC data may become more important in

establishing other benchmark metrics for use and comparisons in future endocrine disrupting chemical screening. This is in effect conducting QA/QC with a purpose since one is not only assuring the quality of these data, but also in enhancing the ability to better differentiate invertebrate endocrine disrupting chemical effects with specific invertebrate MOA effects. This is particularly important in helping better distinguish what different combinations of test endpoint effects mean, given the indirect nature of discerning endocrine effects with these Tier 2 assays. Since mysid tests have been conducted on IGRs which have known AOPs for invertebrate mediated endocrine effects, it would be important for EPA to statistically summarize those test endpoint responses for the MOAs and then use those as a point of comparison with test endpoints for future compounds tested, using appropriate statistical comparisons (e.g. Similarity Analysis) to given probability based comparisons. This would provide a statistical based approach to the tool box for future “weight of evidence” analysis.

3) Importance of defining major/minor deviations in the laboratories performing the assay.

The EPA endocrine disrupting chemical Screening Program would greatly benefit from a Hazard Analysis Critical Control Point (HACCP) approach used by other agencies for other regulatory purposes in reviewing the current crustacean endocrine disrupting chemical bioassays. The performance criteria should identify those test endpoints deviations that represent major and minor non-conformities. Major non-conformities should be significant test deviations of the protocols that may affect bioassay outcome or performance failures on positive and/or negative controls requiring the assay to be re-run. Minor non-conformities should be evaluated to determine if these deficiencies are personnel, facility, equipment, or test support (e.g. suppliers) related. Review of these minor deficiencies should address how they will be corrected and show they should be reported in subsequent bioassays to demonstrate that these inadequacies have been addressed and corrected.

*Question 3. The test protocol should be sufficiently detailed and should include a description of what is measured and how it is measured. The selection of endpoints within the assay should be reflective of the biological processes of interest and the endpoints should be intrinsically relevant and have established sensitivity. The test protocol should demonstrate the ability to measure the endpoints and provide adequate performance criteria for evaluation. **Please comment on the selection, optimization and demonstration of the assay endpoints, as outlined in Sections 3 and 4 of the MTTT ISR.***

Panel Response

The optimization and demonstration of the assay endpoints appears sufficient. However, the amount of inter-laboratory variability and general disagreement in which endpoints are affected and at what dose, suggests that these tests are not yet reliable and repeatable. This suggests that further optimization and demonstration of the mysid and copepod assays is required.

During the course of two life cycles (~60 days) over 20 endpoints are collected with optional biochemical measurements suggested. In general, the test protocol for the MTTT is described in sufficient detail and the requirements for test acceptability presented. However, additional details on animal husbandry and maintenance should be provided as the mysids can be easily disturbed during testing by lighting and other factors. For example, including waterproof curtains around the racks or water tables during the test may help in maintaining proper lighting regimes and reduce unintended stress from unrelated human activities. The endpoints included in the assay are sensitive to and can be reflective of direct adverse effects to the crustacean endocrine system. For example, the test endpoints are hormone regulated endpoints including survival, growth, development and reproduction. Growth and development measurements are based upon measures of time and length and weight measures. Reproduction outputs are based upon survival, fecundity/fertility and number of offspring and timing of reproduction/development. However, the endpoints are also reflective of general toxicity that might be caused by non-crustacean endocrine disruptors. Therefore, the selected endpoints are appropriate for detecting the apical adverse effects of regulatory concern that vertebrate EAT AOP active chemicals may have on invertebrates, regardless of whether or not they disrupt the crustacean endocrine system.

In the absence of direct measurements of invertebrate endocrine disrupting chemical effects (e.g., changes in endogenous ecdysteroid or juvenile hormones, formation of ultraspiracle (USP) ecdysteroid receptor (EcR) complexes or EcR/retinoid-X-receptor (RXR) complexes), only indirect endocrine disrupting chemical specific test endpoints can be inferred. None of the mysid or copepod optimization or demonstration tests identified specific invertebrate endocrine effects *per se*. As is usually the case, single tests cannot be created that are capable of providing evidence for both a mechanism of toxicity and the population level effects of a compound's toxicity. For several of the other Tier 2 assays that are being considered, there is a Tier 1 assay that first indicates a specific endocrine disrupting MOA before the Tier 2 assay is conducted. The Panel recommended inclusion of assays for additional molecular endpoints, such as activation of the ecdysteroid receptor, which have recently been developed as was mentioned in our earlier response to question 2. These endpoints would provide more mechanistic information on effects to be used in interpreting apical effects seen in the MTTT.

The Panel expressed some concern for the loss of highly sensitive subpopulations of males and females in the F0 population during the mysid or copepod multigenerational tests, such a loss may result in having an F1 population that is less sensitive to a chemical and subsequently result in a lower reproductive impact than seen in the F0 population. The Panel speculated that loss of sensitivity may be due to mysids or copepods with increased tolerance or resistance to the tested chemical. A public commenter, speaking on behalf of the Endocrine Policy Forum voiced concern for the higher NOEC in the F2 population vs. the F1 population observed in the mysid test (see public comments submitted to the public e-docket, <http://www.regulations.gov>, docket:EPA-HQ-OPP-2013-0182). Copepod bioassays have also demonstrated increased tolerance and development of resistance to some pesticides, e.g., chlorpyrifos (A. Green, Unpublished doctoral dissertation, 1995). The Panel encouraged the EPA to use information on the development of pesticide resistance to inform the conduct of mysid or copepod Tier 2 tests.

The Panel questioned whether all of the endpoints are needed to characterize the effect of endocrine disrupting chemicals on invertebrates. For example, weight and length measurements taken at two time points rather than at three time points and using first brood size rather than measuring the size of the first two broods may be sufficient.

The Panel indicated that many of the endpoints are interrelated. Given that none of them are definitively linked to endocrine disruption, the integration of these endpoints is critical to understanding cumulative effects of a potential endocrine disrupting chemical. More importantly, inconsistencies among endpoints can be examined for indication of other MOAs. However, the ISR did not characterize the interrelatedness (or lack thereof) of the responses. This issue is noted for endpoints measured in both generations during the assay.

Question 4.1. Demonstration of the test method performance should be based on the testing of reference chemicals representative of the types of substances for which the test method will be used. Test substances should adequately represent an appropriate range of responses and physical/chemical properties for which the test method is proposed to be appropriate. The selection of the most appropriate statistical approaches depends in part on the nature of the data and also on the design of the validation study. Statistical and non-statistical methods used to analyze should be described. Please comment on the selection of test substances and methods (analytical and statistical where appropriate) chosen for the demonstration and validation of the MTTT assay.

Panel Response

The Panel recommended that the choice of experimental design and statistical methods used in demonstration and validation of the MTTT assay be described in greater detail. This recommendation is common to all four Tier 2 ecotoxicity tests reviewed by the Panel.

The Panel reviewed Battelle's re-analysis of the data from the various laboratories because it had the most consistent and clearest description of the methodology used to analyze the inter-laboratory data (see description in the following document, EPA-HQ-OPP-2013-0182-0060.pdf, located in the public e-docket, <http://www.regulations.gov>, docket: EPA-HQ-OPP-2013-1082). The Panel made the following comments and recommendations concerning the statistical analysis.

- 1) The experimental unit on which a particular endpoint is measured should be more clearly described. The Panel stated that the random effect of a replicate is not included in the ANOVA modeling when the data are defined at the unit of a breeding pair or individual mysid (see p. 5). Hence a source of variation is ignored.
- 2) The Panel noted that a two-way mixed model that includes both generations and random effects can be used for more than testing generational differences. It can also be used to perform the tests against controls within each generation, as well as averaged over both generations if there is not a generation treatment interaction. Running one-way ANOVAs for each generation when there is no interaction is redundant and can lead to a loss of power if the assumption of homogeneous variance across generations is valid.

- 3) Assessing normality and heterogeneous variance for such small sample sizes is difficult as the power of the tests is quite low (Razali & Wah, 2011). The Panel pointed out that the EPA's strong emphasis on assessing normality and heterogeneous variance, as well as transformation of the data, is inappropriate given the small sample sizes. The Panel recommended alternative distributions, such as an exponential or gamma, should be considered and guided by studies of the same endpoints in other taxa or for other test substances.
- 4) The use of different transformations and methods for each of the datasets (F0, F1 and combined) defeats the very purpose of model-based inference which assumes that there is a true underlying probability distribution for each endpoint. Transformations and rank-based nonparametric tests arose because there were no statistical methods for non-normally distributed data; i.e., as a convenience to allow analyses to occur given the computing capability available in the past. Unless it is valid to assume that an endpoint measured in the F0 generation has a probability distribution that differs from that for the same endpoint measured in F1 generation, a better approach would be to combine the data and try to identify the true distribution or a close approximation rather than using methods that essentially modify the data to fit the older statistical methods.
- 5) Removal of outliers is inappropriate unless it can be shown that the data are in fact in error, such as a recording mistake or measurements from a malfunctioning machine or piece of equipment. Instead, one should consider alternative methods of analysis that are robust to outliers if in fact there are significant deviations among the numbers in the data. An example of such a method might be to consider using a t-distribution rather than the normal distribution for the data model. This provides heavier tails which may lower in impact of the outliers. Another approach would be to use a robust ANOVA approach such as M-estimation (Huber, 1981) or something similar.
- 6) The ANOVA models should be modified to allow incorporating gender as a fixed effect in the model and its interaction with treatment. This allows for testing whether the endpoints for each gender differ in their response to the treatments and if not, collapsing the data over the two levels for analyses. Not doing this because the statistical package does not provide the analysis is inappropriate; the package should be modified. The methods state that if the test for normality failed, then equality of variances was not tested and the default is use non-parametric approaches. The problem is that the main test, the Kruskal-Wallis test, does make assumptions about shape and variance, namely that every treatment level has the same variance and shape. Hence, this is inappropriate. If the assumption of normality of the residuals is concluded to not be false, then the tests for trend using linear and quadratic contrasts should not be done using non-parametric techniques because of convenience.
- 7) The use of permutation tests for small sample sizes with large number of ties appears to be appropriate but when one has small sample sizes with ties, the tests of normality and

homogeneity would not be of any use. As a result, the Panel was unclear what advantage there is to using a permutation test.

- 8) Mortality and other proportions should be analyzed as binomial random variables with a generalized linear model (one-way ANOVA with a binomial response), not as approximately normal after transformation. The arcsine square root transformation ignores the number of individuals that were observed and treats what is a categorical variable as continuous. This transformation is best suited to data with large number of observations (more than 10 juveniles) and percentages near 0.5. It is not a good transformation when the percentages are near 0 or 100.
- 9) Several endpoint variables are counts, such as the number of young/brood/female. When these values have a small number of possible outcomes, the variable should be analyzed as count data and use distributions such as the Poisson or Negative Binomial distributions. Checks for appropriate distribution choice are available and should also be included in the analysis just as tests for normality or homogeneous variance are used for continuous data. Given the extensive historical literature on the effect of toxins on mysids, it would seem to be advantageous to review that literature for perspective on the distribution of some of the endpoints under study. A meta-analysis would provide much needed background that could aid in better approaches for statistically analyzing the data.

*Question 4.2. The proposed MTTT study design uses three replicates per control and treatment group and additional replicates require greater resources. **Please comment on the proposed level of replication and the subsequent statistical analysis of the data.***

Panel Response

Battelle, one of the EPA's contract laboratories, performed a simple power analysis for estimating the size of the difference that could be observed at 80% power and type I error rate of 0.05 with the current sample sizes and estimated within study variances observed in the validation studies performed for the MTTT (Note: Battelle's analysis is found in the public e-docket, <http://www.regulations.gov>, docket - EPA-HQ-OPP-2013-0182, document number: EPA-HQ-OPP-2013-1082-0047.pdf). This analysis treated the observations as independent with equal variances and sample sizes. As shown in Battelle's analysis, the current number of tanks (or baskets) and the current number of mysids/tank or number of breeding pairs is too small to discern differences, if they exist. Further evidence of this is the lack of comparability between laboratories performing the same test protocol. The Panel recommended that a different power analysis be performed that identifies a sample size that best combines precision and power to discern differences. Further, this should be done in concert with identifying better levels of the test substance concentrations and possibly modifications to the entire approach for statistical analysis. If, for example, the study is looking for a monotonic trend, then six levels are not required. Conversely, six levels are likely needed if the proposed action is either a threshold effect or non-linear with a maximum at intermediate levels. Increasing the number of tanks

and/or individuals per tank and reducing the number of levels of the test substance can provide more precise information at a cost similar to that for the current bioassays.

Outside of the replication problems (only three replicates used), the Panel identified some noticeable, yet understandable, experimental design issues concerning the numbers of animals available for the F1 pairing and reproduction portions of the assay. The trade-off between mortality and loss of statistical power in the F1 generation endpoints in measuring the effects of the toxicant is an enduring issue among toxicologists studying endocrine disruptors. During the initial demonstration and optimization studies, the Panel cited problems with the number of F1 available following reproduction in the F0. The number of surviving F1 individuals is obviously affected by the toxicity and chosen concentration ranges of the toxicant. The Panel noted that for most toxicants the highest exposure concentration was determined by taking 1/3 to 1/100 of the 21-day LC₅₀. This is a difficult decision to make, and must be done cautiously because it will have significant impact on the outcomes of the assay, especially in regard to the F1 generation endpoints. Higher concentrations could be used as there was really low mortality, less than 12% or a LC₂₀ value, that would increase the likelihood of measuring an effect while preventing losses to F1. The MTTT ISR (p. 43) states:

“The usefulness of a 96-hour LC50 percentage (i.e., one-fifth) to select the highest definitive exposure concentration is questionable and may need to be further supported by a longer-term range-finding test (e.g., 14 to 28 days). Setting a level of acceptable mortality after 14 days of exposure at 20%, for example, and estimating the associated LC20 as the maximum dose in the definitive test may be a better procedure for determining the exposure series. This would not guarantee but should increase the likelihood of observing greater, and biologically relevant, percentage differences between means in the two-generation testing scenarios.”

The target number of paired adults for each F1 replicate was seven, but often the common number was N=3, and ranged from 0 to 8. Again, this speaks to the selected compound concentrations and numbers of F0 animals chosen for the assay. As seen in the *Inter-Laboratory Statistical Analysis Report for Mysid 2-Generation Data*, although the median CVs are about equal or larger between the F0 and F1 generations, the F1 CVs tend to exceed the F0 CVs for the same parameter. This is reflected in 15 of 25 F1 CVs exceeding the corresponding F0 CVs. Loss of statistical power would be seen in the F1 generation as a consequence.

The Panel recommended that a threshold number of replicates are needed, i.e., minimal value of replicates, to continue with the F1 exposures regime. One Panel member suggested that the number of replicates might be increased by lowering the number of the individuals within each tank and adding more tanks (which will not add more total individuals). There is an interrelationship between numbers of replicates and numbers of organisms per replicate and the combined influence on statistical power. This relationship will vary with species and test design (among other factors). However, given the current data sets for this assay, it should be possible to investigate this relationship and optimize these two factors. The Panel suggested some further analysis of this question using existing data to generate random generated data sets with larger replicate size to examine the effect of sample size on power so that sample size can be optimized. The Panel emphasized the importance of developing a statistical

manual that addresses statistical analysis methods, approaches and procedures to follow when conducting all four Tier 2 ecotoxicity tests, including the MTTT.

*Question 5. Considering the variability inherent in biological and chemical test methods, a test method needs to be repeatable and reproducible. A test is robust and reliable if the results are repeatable and reproducible within a laboratory and between different laboratories, respectively. A test protocol should provide sufficient guidance to ensure proper and consistent performance across labs and chemicals. **Please comment on the test method robustness and reliability and the repeatability and reproducibility of the results obtained with the MTTT assay.***

Panel Response

1) Statistical comments

The Panel indicated that one laboratory was unable to provide sufficient data for the F1 generation and was excluded from any analyses. Further, the two laboratories that did provide data had results that led to different conclusions. Overall this indicates that either 1) the sample sizes are simply too small for reproducible analyses, 2) there are no treatment effects on the measured endpoints and all significant results were type I errors, 3) the appropriate concentrations of chemicals were not used to show effects, or 4) the laboratories failed to observe and follow the protocols as directed. It is unclear whether the statistical methods are at fault since there is not sufficient information to determine whether the tests for checking assumptions were appropriate or adequate and how the decision for 3 replicates was made. The description given for power determination in the MTTT ISR (p. 43) is insufficient to determine how the final sample size was decided since it does not discuss whether the measures of time to first brood or length (presumably body length) were for individual breeding pairs (time), mysid shrimp (length), or the average of the tank from which they were taken. Since much of the analyses are for experimental units as tank or basket within tank, the determination of power at the individual animal level is non-informative.

2) Non-statistical comments

Validation is a scientific process by which the reliability and relevance of an assay method are evaluated for the purpose of supporting a specific use (ICCVAM, 1996). Reliability is the reproducibility of results from an assay within and between laboratories. Robustness has been defined by other panelists addressing other species as “a system that can handle variability but maintain status quo and consistency” and “the ability to produce results with deviations from protocols.”

The pre-validation process included: 1) demonstration of relevance, 2) development of standard optimized protocol and 3) determination of readiness for validation. Validation in multiple laboratories and demonstration of reliability across laboratories is the final step in this process. Earlier inter-laboratory trials assessing reproducibility of acute toxicity for certain aquatic test species has demonstrated that acute 96 hour LC₅₀ values may vary by a factor of 2.2 to 12 for detergents and a factor of 4.9 (guthion) to 5 (endosulfan) for pesticides among inter-laboratory acute toxicity testing with the same species (Sprague, 1985). Sprague (1985) further reported that in comparing laboratory testing

variance within the same laboratory, that LC50's varied by a factor of 1.14 (ammonia) to 5.5 (copper) for acute toxicity testing in fish. More complex multigenerational testing protocols do not have similar metrics to compare to, but appear to fall within a similar range reported for pesticides, less than a factor of 6 (based on lindane metrics reported by the EPA), despite the increased exposure duration and increased complexities of the assays.

The Panel agreed with EPA's conclusion that many of the test methods need improvement. The Panel agreed with EPA's recommendations to increase the feeding regiment, number of F0 animals at the onset of the assay and removal of the self-starting siphons (see pp. 38-43, section 4.1.3 of the MTTT ISR). An evolution of some endpoints occurred over the course of the studies which allowed for more concise measurement, but disallowed direct comparisons between tests. The MTTT ISR states the following (see p. 42):

“During the optimization study, total young was evaluated as number per female, but the total was calculated two ways. For the first calculation, the reproduction data for all females were treated equally (i.e., reproduction data for females that die before producing a second brood or females not producing a second brood will not be omitted). For the second, only the reproduction data for females that produced a second brood were included. Both sets of calculations were evaluated statistically and the results occasionally differed. The guideline should clarify which approach meets the goals of the testing so that the endpoint will be evaluated consistently by various testing laboratories.”

The Panel agreed that pairs of surviving mysids that do not produce a second brood should not be included in this count, but rather should be included in the “% females reproducing” endpoint. However, since this is not a sensitive endpoint and the numbers of mating pairs were very low in many of the studies, the Panel recommended that the EPA assess the loss of power in the second endpoint.

Question 6.1. The test protocol should be descriptive enough to be fully transferable to a functional laboratory. The protocol should describe the methodology of the assay in a clear and concise manner so that a laboratory could comprehend the objective, conduct the assay, observe and measure prescribed endpoints, compile and prepare data for statistical analyses, and report the results. Section 5 of the MTTT ISR outlines the process of and challenges experienced in the inter-laboratory studies. Please comment on the transferability across labs and provide any suggestions or recommendations for improvement of the MTTT assay.

Panel Response

The Panel indicated that both assays were not yet transferable across laboratories. There were issues of poor compliance with all test protocols and some inconsistencies in the data for growth, development and reproductive endpoints among different laboratories conducting the mysid test, that the Panel envisioned could be corrected by upgrades to the current test protocols and by additional training. The copepod test was not a familiar test for the testing laboratories, who had problems achieving test concentrations in the small dosing chambers of the 96-well plate and who were less

familiar with the life history stages. These problems could be corrected by upgrades to the current test protocols for improving dosing by using larger size dose chambers (e.g., 24-well plate) and by additional taxonomy training. One major issue for both tests is lack of information concerning what is known about the invertebrate endocrine response, given that most test endpoints are indirect assessment endpoints for invertebrate endocrine disrupting chemicals.

The Panel identified two areas that need to be addressed to improve the transferability of the current MTTT bioassays.

Use of Positive Controls. The Panel recommended that future inter-laboratory calibration exercises include use of positive controls with known invertebrate endocrine disrupting chemical effects be run at concentrations that should always illicit these responses. Positive controls will add cost, but are needed to assure that contract laboratories are adequately performing the assays. For example, methoprene or other known invertebrate endocrine disrupting chemicals would be run initially at multiples doses, but after many runs as confidence increases in the results, only a single dose of the positive control would be needed for each new compound tested.

As more and more data are generated for specific invertebrate endocrine disrupting chemical pathways represented by these positive controls, these QA/QC data could be used to establish benchmark metrics and to compare results from future chemicals screened for endocrine disrupting effects. Use of positive controls as part of QA/QC efforts serves two purposes: 1) assuring the quality of these data and 2) enhancing the ability to differentiate invertebrate endocrine disrupting chemical effects with specific invertebrate MOA AOPs. This is particularly important in helping to better distinguish what the different combinations of test endpoint effects actually mean biologically, given the more indirect nature of discerning endocrine effects with these proposed Tier 2 assays.

The Panel recommended that a statistical summary of the results for mysid tests that have been conducted for IGRs, which have known invertebrate-mediated endocrine effects. This collection of responses could be used as a reference for future compounds tested, e.g., similarity analysis for probability based comparisons. This type of analysis would provide a statistical based approach to add to the toolbox for future “weight-of-evidence” analysis.

Importance of Defining Major/Minor Deviations in the Laboratories Performing the Assay.

The Panel indicated that the EPA EDSP would greatly benefit from a Hazard Analysis Critical Control Point (HACCP) approach used by other agencies for other regulatory purposes in reviewing the current crustacean endocrine disrupting chemical bioassays. EPA in writing the performance criteria should identify those test endpoints deviations which represent major and minor non-conformities. Identification of major and minor non-conformities should focus on the conformance with identified control test endpoint criteria in terms of the survival, growth, development and reproductive endpoints. Defining these major and minor non conformities can either be statistically based (e.g. +/- 1 SD of mean) or percentage based (e.g., minor = $\leq 10\%$ of mean and major $> 10\%$ of mean, for a given test endpoint) for the endpoint. Major non-conformities should be significant test deviations of the protocols that may affect bioassay outcome or performance failures on positive and/or negative controls requiring the assay to be re-run. Minor non-conformities should be evaluated to determine if these deficiencies are

personnel, facility, equipment, or test support (e.g., suppliers) related. Review of these minor deficiencies should address how these deficiencies will be corrected and should be reported in subsequent bioassays to demonstrate that these inadequacies have been addressed and corrected.

The Panel stated there were several issues with the description of the experimental design and statistical methods in the MTTT ISR. For example, there were slight differences in the description of the mysid/copepod test protocol, especially for the number of mysids/basket in a tank (see section 2.6 vs. section 2.9.1 in the MTTT ISR). Some of the statistical guidance provided in the MTTT ISR (p. 39, section 4.1.3) was simply incorrect, e.g., describing random effects in the study design as blocking factors. In fact, adding a random block effect requires a general linear mixed model approach, not a generalized linear model as stated in the text (a generalized linear model is a fixed effects model for non-normally distributed data). The statement that the Kruskal-Wallis test could be used for non-monotonic effects is incorrect. In fact, this test is used as a non-parametric alternative to the omnibus F-test in ANOVA and indicates neither a direction of difference nor which treatment means differ.

The Panel agreed with the recommendations given by the EPA in section 4.1.3 of the MTTT ISR; but provided additional recommended changes that will more completely address all the limitations in the protocol: increase the number of organisms used to initiate the test, provide training in how to run the assay and add a positive control that is a known invertebrate endocrine disrupting chemical, e.g., methoprene (see response to charge question 7). Sample sizes of three are much too small to conduct most of the statistical tests being applied in the subsequent analyses.

Panel members were concerned that the statistical methods are described, almost as an afterthought, rather than with the level of detail afforded to any of the laboratory methods. The Panel recommended performing quality control and review of the data prior to statistical analyses and further recommended that modern statistical methods should be used.

Question 6.2. Based on the validation results using the two different invertebrate species, mysid (MTTT) and copepod (HCDRT), the EPA is proposing the mysid protocol as the preferred Tier 2 assay. Please comment on the rationale to recommend the mysid protocol as the preferred Tier 2 invertebrate assay, as described in Section 6.4 of the MTTT ISR.

Panel Response

The Panel had no general concerns with the performance of the mysid protocol; however, the large number of endpoints seems to create duplication of effects; therefore, a reduction in number of endpoints assessed could streamline the protocol without losing power to assess adverse impacts of the test compounds. In their presentation, the EPA acknowledged that there were a few endpoints that had a signal to noise ratio >2 and many endpoints never demonstrated significant effects. (Note: The EPA presentation is found in the public e-docket, <http://www.regulations.gov>, docket: EPA-HQ-OPP-2013-0182.) The Panel suggested an assessment of the endpoints be made to select those endpoints with statistically significant effects and eliminate other endpoints. As an aside, the Panel indicated that when there is no second brood, “0’s” should be reported for the endpoint “2nd broods of females,” rather than creating a metric, “% of females producing 2nd brood”.

The EPA and industry were consistent in their comments regarding limitations in the copepod assay. The copepod assay uses a 96-well plate that is designed for only a few milliliters of test media, which limits the ability of the assay to measure chemical concentrations following exposure and compare them to nominal concentrations. The Panel suggested that this issue could be resolved simply by increasing the size of the wells in the plates slightly, e.g., a 24-well plate, will allow greater media volume to be tested. The Panel stated that a reported lack of professional experience or history working with this assay should not be the most important factor as this assay is very sensitive and will save significant amounts of time, funds and human resources.

The major shortcomings of the copepod assay seem to be more an issue of defining chemical exposures for low volume doses, lack of stage specific morphological character recognition and lack of familiarity with the testing protocol. Two of the issues can be addressed through training and additional experience with running the bioassay. The third can be resolved by decreasing the number of wells in the testing plate from 96-wells to 24-wells, which will increase the test media volume used to conduct the assay. Strengths of the mysid test include the familiarity and experience of laboratories running the bioassay and the large mysid testing database developed by the EPA with data from conventional mysid assays. The proposed Tier 2 mysid multigenerational bioassay does not take advantage of the current extensive mysid database and methods for using these data, such as population modeling.

If the EPA selects the mysid test over the copepod test then Panel recommended additional testing and analysis to support this decision.

- 1) Additional testing of known invertebrate endocrine disrupting chemical compounds, such as methoprene and other known invertebrate hormone affecting chemicals in mysids and copepods, should be conducted to better compare sensitivity and effectiveness of multigenerational effects. This testing would not need to be conducted by multiple laboratories. The goal of the testing is to characterize the sensitivity of invertebrates to a full range of known invertebrate endocrine disrupting chemicals.
- 2) EPA should select chemicals for this testing where existing mysid one generational data exists and the results should be modeled for multigenerational effects and compared with the bioassay results from the additional multigenerational mysid and copepod testing. This will allow two things to be determined: (1) The value of the mysid and copepod multigenerational test in assessing invertebrate endocrine disrupting chemical effects and (2) The value of the current one generational mysid test in predicting known invertebrate endocrine disrupting chemical effects. The Panel questioned the bioassay's ability to discern known invertebrate endocrine disrupting chemical effects across a broad spectrum of invertebrate hormonal effects. Additional data should be collected for chemical exposures to a large variety of chemicals and analyzed. There will be more certainty in which chemicals cause endocrine disrupting chemical effects in mysids/copepods and which assays can discern these effects. Once this established, it will be easier to choose which Tier 2 assay is the best. Inclusion of additional analysis of the

current one generational mysid bioassay results is a clear indication of the Panel's recognition of the extensive information that exists for this bioassay and its proven performance in protecting invertebrates from known invertebrate endocrine disrupting chemicals, e.g., IGRs.

- 3) EPA should perform a meta-analysis to identify which subset of endpoints are most responsive to toxic effects and further what probability distributions are appropriate for those endpoints.

Conclusion

The Panel recommended inclusion of both the mysid and copepod bioassays rather than selecting one bioassay. Both assays have their strengths and weaknesses and are generally equally sensitive in comparisons provided by the EPA in the MTTT ISR. The Panel agreed with the OECD recommendation that both bioassays (classified by OECD as Level 5 tests) be used to provide additional data to enhance invertebrate protection. The OECD uses a broader toolbox approach for testing invertebrates that recognizes the broad and diverse biology of invertebrate species (especially important are the crustaceans, mollusk, annelids, etc.). The Panel stated that EPA's decision to use crustaceans in Tier 2 tests is a wise choice because of the overall greater sensitivity generally seen in crustaceans compared to other species in conventional acute and chronic FIFRA toxicity testing of pesticides and other chemicals. However, the choice of the mysid over the copepod assay at this point seems premature based on the results presented to date. Both assays have similar chemical sensitivities even at low concentrations.

Question 7. The purpose of the validation process is to determine the readiness of a test for inclusion in a testing program. A component of readiness of a test is the evaluation of the usefulness and limitations of the test, including the classes and types of test substances that can and cannot be tested. Please comment on the strengths and/or limitations of the assay, as described in Section 6 of the MTTT ISR.

Panel Response

The Panel summarized the strengths and limitations of the MTTT assay and provided recommendations to address these limitations and strengthen the assay's performance.

Strengths. The Panel summarized the strengths of the MTTT assay as follows:

- 1) A variety of test endpoints are used to assess alterations in growth, development and reproduction which are integrated to predict endocrine disrupting chemical effects.
- 2) There is historical use of the organism in bioassays and extensive experience of testing laboratories with animal husbandry associated with this organism.
- 3) The test is a standardized EPA assay that has been widely used by EPA and other laboratories for assessment of the toxicity of a variety of compounds and the modification of the existing time is simply an additional generational extension of the current assay, which adds maternal transfer as an additional route of exposure.

- 4) The test endpoints are generally simple and easy to measure.
- 5) The assay includes endpoints that are relevant to population-level effects and the results are amenable to common population modeling approaches.
- 6) Males and females are distinguishable by conventional microscopy based on anatomical differences.
- 7) The F0 > F1 metrics were more, or as sensitive, as the F2 endpoints in 8 of 11 chemicals tested (73% of the time). This underscores the importance of the current mysid testing used by EPA for pesticide registration as an important source of information and this view was indicated by public comments provided to the Panel.
- 8) The addition of the second generation assessment for the mysid test provided an additional level of enhanced environmental assessment as 3 of the 11 chemicals (27%) were more sensitive in the F2 generation, suggesting in part, the importance of additional maternal transfer exposure in assessing invertebrate endocrine disrupting chemical effects.
- 9) Test is recommended as part of the OECD testing protocol (Level 5).

Limitations. The Panel summarized the limitations of the assay as follows:

- 1) The large number of endpoints included in the test protocol makes the test tedious. The apparent redundancy among endpoints suggests that they may not have been selected strategically.
- 2) The variability among laboratories in the ring test is troublesome. This variability includes inability of some laboratories to achieve control performance metrics suggesting that testing laboratories are not proficient in conducting this test and/or that the test endpoints are highly variable. Earlier studies with mysids and other invertebrate species have indicated it is not possible to predict life sensitivity in the F1 generation. Similarly, the MTTT cannot predict generational sensitivity among life stage endpoints. This variability also includes endpoints between replicates, treatments, and laboratories. Whether this variability is intrinsic (e.g., animals, feeding, feed quality) or is due to test design or lack of laboratory proficiency is uncertain, but it makes interpretation of the existing data speculative at best. Further, given this variability, it is uncertain if laboratories repeating the bioassay would get the same results.
- 3) Given the variability discussed above there is insufficient sample size to adequately observe significant differences among treatments.
- 4) The addition of a second generation significantly adds to the time and resources required to perform the assay.
- 5) Effects in the first generation may result in an inability to complete the second generation due to insufficient young.
- 6) Some methods were poorly described and may have led to poor test laboratory performance.
- 7) Concern that toxic effects in the F0 generation will result in a less sensitive F1 generation, leading to less sensitivity in F2 metrics. Some pesticides are known to cause pesticide resistance with increased generational exposure in insects.
- 8) The flow through nature of the test generates large volumes of wastes (6000-8000 gallons/test) that increase the test costs (waste disposal of 120-160 55-gallon drums at ~\$1000/drum).
- 9) There is no obvious mechanism from existing Tier 1 studies for triggering the need for this assay. Hence, in the absence of data that allows for discrimination, it is likely that this assay

would always be requested or, alternatively, never be requested. Either of these outcomes would be problematic.

EPA Proposed Refinements. The Panel summarized the EPA’s proposed refinements to the MTTT (p. 69, MTTT ISR) as follows:

- 1) Better method for selecting doses based on results of a 21-day screening test;
- 2) More judicious selection of test chemicals for the validation study based on known arthropod endocrine activity, and
- 3) More refined endpoints to potentially reduce animal use or improve statistical power.

Panel Recommendations. The Panel agreed with these additions, but recommended additional changes that will more completely address all the limitations in the protocol as shown below.

- 1) Increase the number of organisms used to initiate the test to increase probability of survival and adequate reproduction.
- 2) Conduct additional performance studies to increase confidence that each laboratory can conduct all assays according to established protocols and under GLP.
- 3) Add a positive control that is a known invertebrate endocrine disrupting chemicals, e.g. methoprene, with a described invertebrate hormonal pathway effect for inclusion in future testing. A positive control will help establish growth, development and reproductive “fingerprints” that may be useful in better discerning invertebrate endocrine disrupting chemical effects in screening new compounds.

Question 8. There is sufficient evidence to indicate that endocrine disrupting chemicals can disrupt normal development and reproductive success, however the sensitivity of the F1 generation compared to the F0 is less clearly defined. Please comment from a scientific and risk assessment perspective on the value added of multiple generations in the MTTT assay.

Panel Response

The use of multiple generations is intended to capture effects that may not be apparent through exposure of a single generation. As demonstrated by comparison of the mysid sensitivity to chemical exposure by generation (Table 4), LOECs for the F1 generation can be more sensitive than LOECs observed in the F0 generation. For example, of the 11 chemicals available for comparison, F0 mysids were more sensitive to two chemicals, fipronil and 4-*t*-octylphenol, than F1 mysids. Mysids were more sensitive to three chemicals in the F1 generation than the F0 generation, 3, 5-dichloropropane (3, 5-DCP), ketoconazole, and vinclozolin), while the F0 and F1 generations were equally sensitive to the other five chemicals.

Table 4. Comparison LOECs for by generation in mysids.

Chemical	F0 LOEC	F1 LOEC
3,5-DCP	32 µg/L (ppb)	13 µg/L (ppb)
Fenoxycarb	2.5 µg/L (ppb)	2.5 µg/L (ppb)
Fipronil	1.0 µg/L (ppb)	40 µg/L (ppb)
Prochloraz	188 µg/L (ppb)	188 µg/L (ppb)
Flutamide	1.8 mg/L (ppm)	na
Ketoconazole	33.7 µg/L (ppb) [repro]	2.15 µg/L (ppb) [repro]
Octylphenol	0.36 µg/L (ppb)	3.9 µg/L (ppb)
Atrazine	> 193 µg/L (ppb)	> 193 µg/L (ppb) ?
PF-C10 (PFDA)	> 33.3 µg/L (ppb)	> 33.3 µg/L (ppb)
Lindane	0.09 µg/L (ppb)	0.09 µg/L (ppb)
Vinclozolin	> 403 µg/L (ppb) [repro] 120 µg/L (ppb) [other]	403 µg/L (ppb) [repro] 25 µg/L (ppb) [other]

Note: Red numbers indicate the lowest LOEC values for that chemical. All LOECs are based on growth with the exception of those labeled [repro] (reproduction endpoint), or other.

Source: Modification of slide 172, EPA combined presentations.

In Table 5, the effects observed in the F1 generation appear to represent reproductive impacts more consistent with endocrine disruption and the EAT AOP, than effects in growth and development observed in the F0 generation. For example, F0 generation effects include effects on growth and development (i.e., time to molt, time to maturation, 7-d length, and 14-d length), while F1 effects included effects principally on reproductive endpoints (i.e., day to first brood, survival, number of offspring, number of reproducing females, number of 2nd brood, and offspring/female). Growth effects did appear to occur in the F0 generation at concentrations equal to or lower than concentrations where reproductive effects are seen in the F1 generation suggesting that these effects could be predictive of F2 effect concentrations that cause reproductive effects. However, for at least one chemical, 3, 5-DCP, results indicated that reproductive effects in the F1 generation occurred at 3-fold lower concentrations than effect time to molt in the F0 generation. Therefore, effects in F0 may not always be predictive of effects in F1. From this analysis, the use of an F1 generation in the MTTT clearly captures effects at lower concentrations of certain chemicals than if only a single generation test were used. In this case, 30% of chemicals caused greater effects in the F1 generation.

Table 5. Summary of the most sensitive endpoints for the F0 and F1 generations in the MTTTs for seven chemicals.

Chemical	Sensitive endpoint	F0 LOEC	F1 LOEC
3,5-DCP	Total offspring/female	200 µg/L (↓)	7.2 µg/L (↓)
	Time to molt	22 µg/L (↑)	> 490 µg/L
Fenoxycarb	14d length	2.3 µg/L (↓)	NA
	Total offspring/female	4.9 µg/L (↓)	2.5 µg/L (↑)
Fipronil	Time to maturation	< 1 ng/L (↑)	>16 ng/L
	7d length	< 1 ng/L (↓)	>16 ng/L
Ketoconazole	7d weight	< 2.15 µg/L (↓) □	>33.7 µg/L
	# 2 nd brood	50 µg/L (↓)	< 2.15 µg/L (↓)
4-tert-Octylphenol	#2 nd brood	41 µg/L (↓)	> 17.9 µg/L
	# reproducing females	41 µg/L (↓)	> 17.9 µg/L
Lindane	14d length	327 ng/L (↓)	-
	Total offspring	-	159 ng/L (↑)
Vinclozolin	14d length	8.2 µg/L (↑)	-
	Termination length	-	24.8 µg/L (↗)
	Survival (day 7)	600 µg/L (↘)	200 µg/L (↓)
	Day to first brood	119 µg/L (↗)	24.8 µg/L (↑)

(↑) Indicates a statistically significant increase compared to control (ANOVA F-test), based on statistical re-analysis of test data (Battelle 2013). (↓) Indicates a statistically significant decrease compared to control (ANOVA F-test), based on statistical re-analysis of test data (Battelle 2013). (↗) Indicates a statistically significant increasing concentration trend (Linear trend test), based on statistical re-analysis of test data (Battelle 2013). (↘) Indicates a statistically significant decreasing concentration trend (Linear trend test), based on statistical re-analysis of test data (Battelle 2013). (-) Indicates an endpoint was measured, but was not statistically significantly different from control. Values in **bold** represent the lower value in the comparison of F0 and F1 LOECs.

Source: Information derived from the Mysid Two-Generation Toxicity Test and Harpacticoid Copepod Development and Reproduction Test Integrated Summary Report.

Table 6 illustrates the different endpoints measured and the effect of sample size in the F1 generation, especially at the termination point. Of the 10 chemicals tested where an effect was observed, the F0 was the most sensitive life stage in 5 out of the 10 chemicals (50%), the F1 was the most sensitive stage in 2 out of the 10 compounds tested (20%) and equivalent sensitivity between F0 and F1 was observed in 3 out of the 10 compounds tested (30%). Reproduction, growth/development and survival were measured as responsive endpoints for both F0 and F1. The addition of multiple generations to the mysid test provides information pertinent to risk assessment that would be lost if the F1 generation was removed from the test (Table 6). In some studies, it was apparent that the number of organisms carried into the F1 generation tests was not sufficient to generate statistically meaningful results especially after loss of animals (see the response to charge question 5 for more details). As a result, the Panel recommended that the F1 generation study be continued; however, the protocol should be amended to use larger sample sizes in the F0 generation so that a larger number of animals are available for continuation of tests in the F1 generation study.

Table 6. Significant effects of compounds tested in mysid demonstration studies included in the MTTT ISR (including both results from Dunn’s and Dunnett’s tests).

Compounds tested	Reproduction		Growth and Development		Survival		Most Sensitive Stage		
	F0	F1	F0	F1	F0	F1	F0	F1	F0=F1
Atrazine #									
3,5-Dichlorophenol	x	x	x					x	
Fenoxycarb*	x	x	x						x
Fipronil*			x	x	x	x	x		
Flutamide*	x	NA	x	NA			x		
Ketoconazole	x	x	x						x
Lindane		x	x					x	
4-tert-Octylphenol*	x		x				x		
Perfluorodecanoic acid			x				x		
Prochloraz*	x	NA	x	NA			x		
Prochloraz		x	x	x					x

* = Low replicate numbers remaining in F1 generation in one or more treatments; [na] = Not assessed due to limited number of offspring; [x] = Effects measured; [empty] = No effects measured. Source: Information derived from the MTTT ISR.

References

- Adkins-Regan E. (2008). Do hormonal control systems produce evolutionary inertia? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1497): 1599-1609.
- Adkins-Regan, E., Banerjee S. B., Correa, S. M., Schweitzer C. (2013). Maternal Effects in Quail and Zebra Finches: Behavior and Hormones. *Gen Comp Endocrinol*. 2013 Sep 1;190: 34-41.
- Angle B. M., Do R. P., Ponzi D., Stahlhut R.W., Drury B. E., *et. al.* (2013). Metabolic disruption in male mice due to fetal exposure to low but not high doses of bisphenol A (BPA): Evidence for effects on body weight, food intake, adipocytes, leptin, adiponectin, insulin and glucose regulation. *Reprod Toxicol*. (2013), <http://dx.doi.org/10.1016/j.reprotox.2013.07.017> [Epub ahead of print].
- ASTM [American Society for Testing and Materials].(1990). Standard Practice for Conducting Reproductive Studies with Avian Species. Vol 11.94. Philadelphia: ASTM. Pp. E1062-E1086. E1086.
- Ben M. G. and Yohai V. J. (2004). Quantile-Quantile Plot for Deviance Residuals in the Generalized Linear Model, *Journal of Computational and Graphical Statistics*. Vol. 13, 36-47.
- Bitman J., Cecil H. C., Harris S. J., Fries G. F. (1969). DDT induces a decrease in eggshell calcium. *Nature*. 224(5214): 44-46.
- Blair R. C., Higgins J. J. (1985). A Comparison of the Power of the Paired Samples Rank Transform Statistic to that of Wilcoxon's Signed Ranks Statistic. *Journal of Educational and Behavioral Statistics* 10(4): 368-383.
- Blair R. C., Sawilowsky S. S., Higgins J. J. (1987). Limitations of the rank transform in factorial ANOVA. *Communications in Statistics: Computations and Simulations B16*: 1133-1145.
- Bradley D. J., Towle H.C., Young W.S., 3rd. (1992). Spatial and temporal expression of alpha- and beta-thyroid hormone receptor mRNAs, including the beta 2-subtype, in the developing mammalian nervous system. *J. Neurosci*. 12:2288-302.
- Brande-Lavridsen N, Christensen-Dalsgaard J, Korsgaard B. (2010). Effects of ethinylestradiol and the fungicide prochloraz on metamorphosis and thyroid gland morphology in *Rana temporaria*. *The Open Zoology Journal* 3: 7-16.
- Bulloch D., Lavado R., Forsgren K., Beni S., Schlenk D., Larive C. (2012). Analytical and Biological Characterization of Halogenated Gemfibrozil Produced through Wastewater Chlorination. *Environmental Science and Technology* 46:5583-5589.
- Bustin S. A., Benes V. Garson J.A., Hellems J., Huggett J., Kubista M., Mueller R., Nolan T., Pfaffl M. W., Shipley G. L., Vandesompele J. and Wittwer C. (2009). The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* 55: 611-622.

Carr J. A., Patiño R. (2011). The hypothalamus–pituitary–thyroid axis in teleosts and amphibians: Endocrine disruption and its consequences to natural populations. *General and Comparative Endocrinology* 170: 299-312.

Chang T. S., Glick B., Winter A. R. (1955). The significance of the bursa of Fabricius of chickens in antibody production. *Poultry Sci.*34:1187.

Clotfelter E. D., Bell A. M., Levering K. R. (2004). The role of animal behaviour in the study of endocrine-disrupting chemicals. *Animal Behaviour*, 68:665-676

Conover W. J., Iman R. L. (1976). On some alternative procedures using ranks for the analysis of experimental designs. *Communications in Statistics A5*: 1349–1368.

Conover W. J., Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician* 35(3): 124–129.

Crofton K. M., Foss J. A., Hass U., Jensen K. F, Levin E. D., Parker S. P. (2008). Undertaking positive control studies as part of developmental neurotoxicity testing: a report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints. *Neurotoxicol Teratol* 30:266-287.

Davison K. L., Sell J. L. (1972). Dieldrin and *p,p'*-DDT effects on egg production and eggshell thickness of chickens. *Bull. Environ. Contam. Toxicol.* 7(1): 9-18.

De Wilde R., Swevers L., Soin T., Christiaens O., Rougé P., Cooreman K., Janssen C. R., Smaghe G. (2013). Cloning and functional analysis of the ecdysteroid receptor complex in the opossum shrimp *Neomysis integer* (Leach, 1814). *Aquatic Toxicology*. Vol. 130-131: 31-40. 10p.

Dean C. E, Hargis B. M, Hargis P. S. (1991). Effects of zinc toxicity on thyroid function and histology in broiler chicks. *Toxicol Lett.* 57(3):309-318.

Distel C. A., Boone M. D. (2011). Insecticide has asymmetric effects on two tadpole species despite priority effects. *Ecotoxicology* 20: 875-884.

Dunn P. K., Smyth G. K. (1996). Randomized Quantile. *Journal of Computational and Graphical Statistics.* 5: 236-244.

Einot I., Gabriel K. R. (1975). A Study of the Powers of Several Methods of Multiple Comparisons. *Journal of the American Statistical Association*, 70: 351.

Eng M. L., Elliott J. E., MacDougall-Shackleton S. A., Letcher R. J., Williams T. D. (2012). Early exposure to 2,2',4,4',5-pentabromodiphenyl ether (BDE-99) affects mating behavior of zebra finches. *Toxicological Sciences* 127:269-276.

Fitzpatrick J., Mendez E., Walls, I. (2008). Introduction to the ILSI Research Foundation/Risk Science Institute reports from the expert working group on neurodevelopmental endpoints. *Neurotoxicology and Teratology*, 30(4): 263-265.

Frésard L., Morisson M., Brun J. M., Collin A., Pain B., Minvielle F., Pitel F. (2013). Epigenetics and phenotypic variability: some interesting insights from birds. *Genetics Selection Evolution*. 45(1): 16.

Fulton M., Key P., Wirth E., Leight A.K., Daugomah J., Bearden D., Sivertsen S., Scott G.I. (2006). An evaluation of contaminated estuarine sites using sediment quality guidelines and ecological assessment methodologies. *Ecotoxicology* 15(7): 573-581.

Gaertner K., Chandler G. T., Quattro J., P. Ferguson L., Sabo-Attwood T. (2012). Identification and expression of the ecdysone receptor in the harpacticoid copepod, *Amphiascus tenuiremis*, in response to fipronil. *Ecotoxicology and Environmental Safety*. 76: 39–45.

Garcia-Munoz E., Guerrero F., Parra G. (2010). Intraspecific and interspecific tolerance to copper sulphate in five Iberian amphibian species at two developmental stages. *Arch Environ Contam Toxicol* 59: 312-321.

Gentles A., Surlis J., Smith E.E. (2005). Evaluation of adult quail and egg production following exposure to perchlorate-treated water. *Env. Tox. Chem.* 24(8):1930–1934.

Glick G., Chang T. S., Jaap R. G. (1956). The bursa of Fabricius and antibody production. *Poultry Sci.*35: 224–234.

Gould J. C., Cooper K. R., Scanes C. G. (1999). Effects of polychlorinated biphenyls on thyroid hormones and liver type I monodeiodinase in the chick embryo. *Ecotoxicol. Environ. Saf.* 43(2):195-203.

Groothuis T. G., Schwabl H. (2008). Hormone-mediated maternal effects in birds: mechanisms matter but what do we know of them? *Philosophical Transactions of the Royal Society B: Biological Sciences*. 363(1497):1647-1661.

Guiguen Y., Fostier A., Piferrer F., Chang C-F. (2010). Ovarian aromatase and estrogens: A pivotal role for gonadal sex differentiation and sex change in fish. *General and Comparative Endocrinology*. 165: 352-366.

Gupta S., Kanungo M.S. (1996). Modulation of vitellogenin II gene by estradiol and progesterone in the Japanese quail. *Biochem. Biophys. Res. Commun.* 222(1):181-185.

Hackett S.J., Kimball R.T., Reddy S., Bowie R.C.K., Braun E.L., Braun M.J., Chojnowski J.L., Cox W.A., Han K.-L., Harshman J., Huddleston C., Marks B.D., Miglia K.J., Moore W.S., Sheldon F.H., Steadman D.W., Witt C.C., Yuri T. (2008). A phylogenomic study of birds reveals their evolutionary history. *Science*. 320: 1763-1768.

Hala D., Huggett D.B., Burggren W.W. (2012). Environmental stressors and the epigenome. *Drug Discovery Today: Technologies*. In Press.

Headrick T. C. (1997). Type I error and power of the rank transform analysis of covariance (ANCOVA) in a 3 x 4 factorial layout. Unpublished doctoral dissertation, University of South Florida.

Hettmansperger T. P., McKean J. W. (1998). Robust nonparametric statistical methods. Kendall's Library of Statistics 5(1st ed.). London: Edward Arnold. pp. xiv+467 pp.

Hickey J. J., Anderson D. W. (1968). Chlorinated hydrocarbons and eggshell changes in raptorial and fish-eating birds. *Science* 162(3850): 271-273.

Higgins J. J., Tashtoush S. (1994). An aligned rank transform test for interaction. *Nonlinear World* 1: 201-211.

Hoenig J. M., Heisey D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *American Statistician*, 55(1): 1- 6.

Holm L., Blomqvist A., Brandt I., Brunström B., Ridderstråle Y. Berg C. 2006. Embryonic exposure to *o,p'*-DDT causes eggshell thinning and altered shell gland carbonic anhydrase expression in the domestic hen. *Environ. Toxicol. Chem.* 25(10): 2787-2793.

Holmes S.B., Boag P.T. (1990). Effects of the organophosphorus pesticide fenitrothion on behaviour and reproduction in zebra finches. *Environmental Research*, 53: 62-75.

Huber P. J. (1981), *Robust Statistics*. New York: John Wiley & Sons

Huggett D.W., Brooks B.W., Peterson B., Foran, C. M., Schlenk D. (2002). Toxicity of select beta-adrenergic receptor blocking pharmaceuticals (β -blockers) on aquatic organisms. *Archives of Environmental Contamination and Toxicology* 43: 229-235.

Hurlburt, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*. 54(2): 187 – 211.

Iman R. L. (1974). A power study of a rank transform for the two-way classification model when interactions may be present. *Canadian Journal of Statistics*. 2(2): 227–239.

Iman R. L., Conover W. J. (1976). A comparison of several rank tests for the two-way layout (SAND76-0631). Albuquerque, NM: Sandia Laboratories.

Iwamatsu T., Nakamura H., Ozato K., Wakamatsu Y. (2003). Normal growth of the “see-through” medaka. *Zool. Sci.* 20: 607–615.

Jayakumar K., Mohan K., Swamy, H. D. N., Shridhar N. B., Bayer M. D. (2010). Study of nephrotoxic potential of acetaminophen in birds. *Toxicol Int.* 17(2): 86–89.

Jetz W., Thomas G. H., Joy J. B., Hartmann K., Mooers A. O. (2012). The global diversity of birds in space and time. *Nature* 491:444-448.

Jones P. D., Hecker M., Wiseman S., Giesy J. P. (2013). Birds. In P. Matthiessen (ed.) *Endocrine Disruptors: Hazard Testing and Assessment Methods*. John Wiley & Sons, Inc. Hoboken, NJ. Pp 272-303.

Junges C. M., Peltzer P. M., Lajmanovich R. C., Attademo A. M., Cabagna Zenklusen M. C., Basso A. (2012). Toxicity of the fungicide trifloxystrobin on tadpoles and its effect on fish-tadpole interaction. *Chemosphere*, 87: 1348-1354.

Kenward M. G., Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53: 983–997.

Kepner, J. L., Wackerly, D. D. (1996). On rank transformation techniques for balanced incomplete repeated-measures designs. *Journal of the American Statistical Association*, 91(436): 1619–1625.

Kolaja, G. J., Hinton, D. E. (1977). Effects of DDT on eggshell quality and calcium adenosine triphosphatase. *J. Toxicol. Environ. Health*, 3(4): 699-704.

Kortenkamp A., Evans R., Martin O., McKinlay R., Orton F., Rosivatz E. (2012). State of the art assessment of endocrine disruptors: Final Report. Annex 1 – Summary of the State of the Science. European Commission, Directorate for the Environment.
(http://ec.europa.eu/environment/endocrine/documents/4_Annex%201%20Summary%20State%20of%20Science%20ED%20V6.pdf).

Kramer C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications, *Biometrics*, 12: 309–310.

Lewis S. S., Weber G. J., Freeman J. L., Sepúlveda, M. S. (2012). Molecular epigenetic changes caused by environmental pollutants. In S. C. Sahu (ed.) *Toxicology and Epigenetics*. John Wiley & Sons, Ltd, Sussex, UK. pp. 73-109.

Magnoli A. P., Monge M. P., Nazar F. N., Magnoli C. E., Cavaglieri L. R., Bagnis G., Dalcero A.M., Marin R. H. (2012). Combined effects of aflatoxin B1 and corticosterone treatment on selected performance indices and liver histopathology in Japanese quail. *Poult Sci.* 91(2):354-361.

Mason R., Tennekes H., Sánchez-Bayo F., Jepsen, P. U. (2013). Immune suppression by neonicotinoid insecticides at the root of global wildlife declines. *Journal of Environmental Immunology and Toxicology*. 1(1)3 -12.

McCormack J. E., Harvey M. G., Faircloth B. C., Crawford N. G., Glenn T. C., *et al.* (2013). A Phylogeny of Birds Based on Over 1,500 Loci Collected by Target Enrichment and High-Throughput Sequencing. *PLoS ONE* 8(1): e54848. doi:10.1371/journal.pone.0054848

McNabb F. M. (2007). The hypothalamic-pituitary-thyroid (HPT) axis in birds and its role in bird development and reproduction. *Crit Rev Toxicol.* 37(1-2): 163-193.

Meyer R. K., Rao M. A., Aspinall R. L. (1959). Inhibition of the development of the bursa of Fabricius in the embryos of the common fowl by 19-nortestosterone. *Endocrinology*. 64: 890–897.

Meylan S., Miles D. B., Clobert J. (2012). Hormonally mediated maternal effects, individual strategy and global change. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 367(1596): 1647-1664.

Miller, D. H. *et al.* (2007). Linkage of biochemical responses to population-level effects: A case study with vitellogenin in the fathead minnow. *Environ. Toxicol. Chem.* 26: 521-527.

Milliken G. A., Johnson, D. E. (2006). *Analysis of Messy Data, Vol. 1: Designed Experiments*, 2nd Edition. Chapman & Hall/CRC Press, Boca Raton, FL.

Mueller A. P., Wolfe H. R., Meyer, R. K. (1960). Precipitin production in chickens. XXI. Antibody production in bursectomized chickens and in chickens injected with 19-nortestosterone on the fifth day of incubation. *J. Immunol.* 85: 172–179.

Munn S., Goumenou M. (2013). Key scientific issues relevant to the identification and characterization of endocrine disrupting substance: Report of the Endocrine Disruptors Expert Advisory Group. European Commission Joint Research Centre, Institute for Health and Consumer Protection, Ispra, Italy.

Nanna M. J. (2002). Hotelling's T² vs. the rank transformation with real Likert data. *Journal of Modern Applied Statistical Methods.* 1: 83–99.

OECD [Organisation for Economic Cooperation and Development]. (1993). *OECD Guidelines for the Testing of Chemicals. Section 2—Effect on Biotic Systems: Test Guideline 206: Avian Reproduction Test (adopted April 1984).* Paris, OECD.

OECD (Organization for Economic Cooperation and Development). (1996). *Final Report of the OECD Workshop on Harmonization of Validation and Acceptance Criteria for Alternative Toxicological Test Methods.* Paris, OECD.

OECD (Organization for Economic Cooperation and Development). (2005). *Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment.* Paris, OECD.

OECD (Organization for Economic Cooperation and Development). (2007). *Detailed Review Paper for Avian Two-Generation Toxicity Test.* Organisation for Economic Co-Operation and Development Series on Testing and Assessment Number 74. Environment Directorate, Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology. ENV/JM/MONO(2007). 21. JT03230947. Available at: [http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono\(2007\)21&doclanguage=en](http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2007)21&doclanguage=en).

OECD (Organization for Economic Cooperation and Development). (2009). *Guideline 28 for the Testing of Chemicals; Proposal for a New Test Guideline 223: Avian Acute Oral Toxicity Test Draft,* November 2009.

OECD. (Organization for Economic Cooperation and Development). (2011). *Avian Two-generation Toxicity Test in the Japanese Quail, OECD Draft Guideline for Testing of Chemicals.* Organisation for Economic Co-Operation and Development.

Ottinger, M. A., Lavoie, E., Thompson N., Barton A., Whitehouse K., Barton M. Viglietti-Panzica C. (2008). Neuroendocrine and behavioral effects of embryonic exposure to endocrine disrupting chemicals in birds. *Brain Research Reviews*. 57(2): 376-385.

Raldua D. Babin P. J. (2009). Simple, rapid zebrafish larva bioassay for assessing the potential of chemical pollutants and drugs to disrupt thyroid gland function. *Environ. Sci. Technol.* 43: 6844-50.

Rattner B. A., Horak K. E., Lazarus R. S., Eisenreich K. M., Meteyer C. U., Volker S. F., Campton C. M., Eisemann J. D., Johnston J. J. (2012). Assessment of toxicity and potential risk of the anticoagulant rodenticide diphacinone using Eastern screech-owls (*Megascops asio*). *Ecotoxicology*. 21(3): 832-846.

Rattner B. A., Horak K. E., Warner S. E., Day D. D., Meteyer C. U., Volker S. F., Eisemann J. D., Johnston J. J. (2011). Acute toxicity, histopathology, and coagulopathy In. American kestrels (*Falco sparverius*) following administration of the rodenticide diphacinone. *Environ. Toxicol. Chem.* 30(5): 1213-1222.

Rao J. N. K., Scott A. J. (1992). A simple method for the analysis of clustered binary data. *Biometrics* 48: 577-585.

Rao J. N. K., Scott A. J. (1999). A simple method for analyzing overdispersion in clustered Poisson data. *Statistics in Medicine*. 18: 1373-1385.

Razali N. M., Wah, Y. B. (2011). Power comparison of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Darling-Anderson tests. *Journal of Statistical Modeling and Analytics*. 2(1): 21-33.

Saita E., Hayama S., Kajigaya H., Yoneda K., Watanabe G., Taya K. (2004). Histologic changes in thyroid glands from great cormorant (*Phalacrocorax carbo*) In. Tokyo Bay, Japan: Possible Association with Environmental Contaminants. *J. Wildl. Dis.* 40(4): 763–768.

SAS Institute. (1985). SAS/stat guide for personal computers (5th ed.). Cary, NC.

SAS Institute. (1987). SAS/stat guide for personal computers (6th ed.). Cary, NC.

SAS Institute. (2008). SAS/STAT 9.2 User's guide: Introduction to Nonparametric Analysis. Cary, NC.

Sawilowsky S. (1985). A comparison of random normal scores test under the F and Chi-square distributions to the 2x2x2 ANOVA test. *Florida Journal of Educational Research*. 27: 83–97.

Sawilowsky S. (1985). Robust and power analysis of the 2x2x2 ANOVA, rank transformation, random normal scores, and expected normal scores transformation tests. Unpublished doctoral dissertation, University of South Florida.

Sawilowsky S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research* 60: 91–126.

Sawilowsky S., Blair R. C., Higgins, J. J. (1989). An investigation of the type I error and power properties of the rank transform procedure in factorial ANOVA. *Journal of Educational Statistics*. 14(3): 255–267.

Schneider S, Kaufmann W., Strauss V., van Ravenzwaay B. (2011). Vinclozolin: A feasibility and sensitivity study of the ILSI-HESI F1-extended one-generation rat reproduction protocol. *Regulatory Toxicology and Pharmacology*. 59: 91-100.

Shibuya K., Wada M., Mizutani M., Sato K., Itabashi M., Sakamoto T. (2005). Vitellogenin detection and chick pathology are useful endpoints to evaluate endocrine-disrupting effects in avian one-generation reproduction study. *Environ. Toxicol. Chem.* 24(7): 1654-1666.

Sonne C., Verreault J., Gabrielsen G. W., Ltcher R. J., Leifson P. S. Iburg T. (2010). Screening of thyroid histology in organohalogen-contaminated glaucous gulls (*Larus hyperboreus*) from the Norwegian Arctic. *Tox. Envir. Chem.* 92: 1705-1713.

Song M., Kim Y-J., Song M-K., Choi H-S., Park Y-K., Ryu J-C. (2011). Identification of classifiers for increase or decrease of thyroid peroxidase activity in the FTC-238/hTPO recombinant cell line. *Environmental Science and Technology* 45: 7906–7914.

Sprague J. B. (1985). Factors that modify toxicity. In. Rand, G. M. and S. R. Petrocelli, *Fundamentals of Aquatic Toxicology*. pp. 126-127.

Starck J. M., Ricklefs, R. E. (Eds.). (1998). *Avian growth and development: evolution within the altricial-precocial spectrum* (No. 8). Oxford University Press.

Stroup W. W. (2013). *Generalized Linear Mixed Models*. Chapman & Hall/CRC Press, Boca Raton, FL.

Suzuki A., Tanaka M., Shibata N. (2004). Expression of aromatase mRNA and effects of aromatase inhibitor during ovarian development in the medaka, *Oryzias latipes*. *J. Exp. Zoolog. Part A. Comp. Exp. Biol.* 2004. May 1; 301(5): 460.

Thompson, G. L. (1991). A note on the rank transform for interactions. *Biometrika* 78(3): 697–701.

Thompson G. L., Ammann L. P. (1989). Efficiencies of the rank-transform in two-way models with no interaction. *Journal of the American Statistical Association*. 4(405): 325–330.

Troan B. V., Boyle, M. H., Hobbie K. R. (2012). Final report. Histopathology Evaluation of Reproductive and Selected Visceral and Brain Tissue from the Avian 2-Generation Test in Vinclozolin Exposed Adult Japanese Quail.

Tukey J.W. (1994). The Problem of Multiple Comparisons, H. I. Braun, ed., In. *The Collected Works of John W. Tukey*, Volume 8, 1994, H. I. Braun (Ed.), Chapman and Hall, New York.

United States Environmental Protection Agency (USEPA). (2012a). *Ecological Effects Test Guidelines: OCSPP 850.2300: Avian Reproduction Test*. EPA-712-C-023. Office of Chemical Safety and Pollution Prevention United States Environmental Protection Agency.

United States Environmental Protection Agency (USEPA). (2012b). Ecological Effects Test Guidelines: OCSPP 850.2200: Avian Dietary Toxicity Test. EPA-712-C-024. Office of Chemical Safety and Pollution Prevention United States Environmental Protection Agency.

vom Saal F. S., Richter C. A., Ruhlen R. R., Nagel S. C., Timms B. G., Welshons, W. V. (2005). The importance of appropriate controls, animal feed and animal models in interpreting results from low-dose studies of bisphenol A. *Birth Defects Research Part A: Clinical and Molecular Teratology*. 73(3): 140-145.

Wang N., Braun E. L., Kimball R. T. (2011). Testing hypotheses about the sister group of the passeriformes using an independent 30-locus data set. *Mol. Biol. Evol.* 29(2): 737-750.

Warren D. F. Clayton H. E., Arthur A. P. *et al.* (2010). The genome of a songbird. *Nature*. 464: 757-762.

Welch B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, 38, 330–336.

Yokota H., Eguchi S., Nakai M. (2011). Development of an in vitro binding assay for ecdysone receptor of mysid shrimp (*Americamysis bahia*). *Aquatic Toxicology*. Vol. 105 Issue 3/4: 708-716. 9p.

Zala S. M., Penn D. J. (2004). Abnormal behaviours induced by chemical pollution: a review of the evidence and new challenges. *Animal Behaviour*. 68: 649-664

Zann R. (1996). *The Zebra Finch: A Synthesis of Field and Laboratory Studies*. Oxford University Press: Oxford, UK.

Zhang X., Ho S. M. (2011). Epigenetics meets endocrinology. *Journal of Molecular Endocrinology*. 46(1): R11-R32.

Appendix 1: Histopathology Scoring

The Panel noted that the quality of the histology was excellent. A harmonized system of diagnostic terms and severity grading was established by pathologists from each study laboratory.

1) Grading scores

For example, “*Phenotypic morphology of male and female gonads in F2 generation embryos was evaluated using the following criteria with completely male gonads (grade 1) and completely female gonads (grade 5) at the extremes and intermediate / mixed phenotypes in the middle (grades 2-4).*”

- *Grade 1: Morphology is consistent with male gonad.*
- *Grade 2: Eighty percent of gonad contains spermatogonia with discrete evidence of ovarian tissue, as few as 1-2 occurrences of oogonia or ovarian cavity. These foci represent 20% or less of the gonad tissue evaluated.*
- *Grade 3: Gonad contains distinct foci of ovarian and testicular tissue that encompasses 50% of the gonad-section evaluated, respectively.*
- *Grade 4: Eighty percent of the gonad contains ovarian tissue. Foci of testicular tissue, spermatogonia, and occasional seminiferous tubules, are present and represent at least 20% of the gonad.*
- *Grade 5: Morphology is consistent with a female gonad with an ovarian cavity and cortex.*

2) Severity Scores

Similarly, a severity scoring system was employed. The severity of a given change was scored as one of five categories:

- *NR < 5% tissue section affected*
- *Minimal 5-30% of cells affected*
- *Mild 30-60% of tissue or cells affected*
- *Moderate 60-80% of tissue or cells affected*
- *Severe Over 80% of tissue or cells affected.*

The Panel commented that it was not clear why morphometric analysis was not used instead of these scores. If the percentage of the tissue affected is known, why not use it? This would facilitate statistical analyses as the data would be continuous.

- 3) Lost samples.** The Panel expressed concern for the low number of replicates, particularly in view of the loss of some samples. As stated in the LAGDA ISR, “*The pineal gland was present in only 2 of 10 FI males and 1 FI females.*” In the experience of one of the Panel members, all birds have pineal glands. These may be attached to the cranium; thus, unless care is taken, they can be lost during dissection.

- 4) Epididymis lesions.** Adverse effects of vinclozolin on the epididymis were consistently observed. These results are consistent with the anti-androgenic effects of this putative endocrine disrupting chemical. There was an increase in the incidence and severity of efferent duct dilation observed in vinclozolin-exposed F0 and F1 adult males from all three laboratories. This was one of the most consistent endpoints. See Appendix 2 for analysis by one panelist.
- 5) Lesions in non-reproductive organs (F1 males and females combined).** There was a lack of consistent effects of vinclozolin on histology for most tissues examined. For instance, the report stated that for samples from Laboratory 1: *“There were no lesions attributed to vinclozolin exposure in the adrenal gland, heart, liver, kidney, brain (pre-optic nucleus-POM, pituitary or pineal gland), and thyroid gland of male or female birds.”* In contrast, there appeared to be a consistent effect of vinclozolin on the liver. For instance, in samples from Laboratory 3 *“minimal to mild, random multifocal vacuolization was observed. Hepatic vacuolization was characterized by random and multifocal hepatocytes.”*

Appendix 2: Summary of Histopathology Data for Vinclozolin from Three Laboratory Studies

1) Summary of epididymal histopathology data

Table 1. Summary of the effect of vinclozolin on epididymal histopathology using samples from laboratory 2 (data from Troan *et al.* 2012)

	F0		F1	
	Vehicle	Vinclozolin (1000 ppm)	Vehicle	Vinclozolin (1000 ppm)
Epididymis				
Unremarkable	2	0	3	0
Duct dilation				
Minimal		2		1
Minor	1	2	1	1
Moderate		1		1
Severe				

If scores are assigned in the following manner:

- 0 = no abnormal histopathology,
- 1 = minimal abnormalities,
- 2 = minor abnormalities,
- 3 = moderate abnormalities, and,
- 4= severe abnormalities,

then it is possible to analyze the mean scores for the histopathology of epididymal tissues from the three laboratories. No abnormal testicular histopathology was observed in birds from F0 and F1 generations. The data on testis histopathology was pooled results from F0 and F1 generations. These data are presented in table 2A-C below. There was a marked consistency in epididymal responses to vinclozolin and in the lack of a testicular response to vinclozolin.

2) Effects of vinclozolin on pathohistology: results from laboratories 1-3

The effect of vinclozolin on pathohistology of the testes, epididymis and bursa Fabricius in samples from laboratories 1- 3 is summarized in Ttables 2A, 2B, and 2C, respectively.

Table 2A. Effect of vinclozolin on the pathohistology of the testes, epididymis and bursa Fabricius in samples from laboratory 1. Data shown as mean scores for lesions in testes, epididymis and bursa (with number of replicates).

	Vehicle	Vinclozolin (1000 ppm)
Testes (F0 +F1)	0 (N=10)	0 (9)
Epididymis		
Duct dilation (F0)	0.7 (3)	1.8 (5)
Duct dilation (F1)	0.5 (4)	2.0 (4)
Efferent Duct Vacuolization (F0)	0.3 (3)	2.0 (4)
Efferent Duct Vacuolization (F1)	0.5 (4)	2.0(4)
Bursa Fabricius	2.2 (4)	1.6 (5)

Table 2B. Effect of vinclozolin on the pathohistology of the testes, epididymis and bursa Fabricius in samples from laboratory 2. Data shown as mean scores for lesions in testes, epididymis and bursa (with number of replicates).

	Vehicle	Vinclozolin (1000 ppm)
Testes (F0 +F1)	0 (N=10)	0 (9)
Epididymis		
Duct dilation (F0)	0.8 (5)	2.2 (4)
Duct dilation (F1)	0 (5)	1.8 (5)
Efferent Duct Vacuolization (F0)	0 (2)	2.7 (3)
Efferent Duct Vacuolization (F1)	0 (5)	1.8 (5)

Table 2C. Effect of vinclozolin on the pathohistology of the testes, epididymis and bursa Fabricius in samples from laboratory 3. Data shown as mean scores for lesions in testes, epididymis and bursa (with number of replicates).

	Vehicle	Vinclozolin (1000 ppm)
Testes (F0 +F1)	0 (N=10)	0 (10)
Epididymis		
Duct dilation (F0)g20	1 (3)	2.2 (4)
Duct dilation (F1)	0.7 (4)	1.0 (3)
Efferent Duct Vacuolization (F0)	0.7 (3)	2.0 (4)
Efferent Duct Vacuolization (F1)	1 (4)	0.5 (1)

3) Combined data from all three laboratories

The data in tables 2A, 2B and 2C were combined to assess the overall effects of vinclozolin on epididymal and testis histopathology. These data are summarized in Table 3. The combined laboratory results indicate that there was no effect of vinclozolin on testicular histopathology. There was a consistent effect of vinclozolin on the epididymis with evidence for both ductal dilation and efferent vacuolization. These data on epididymal responses to vinclozolin are consistent with vinclozolin exerting an anti-androgenic effect.

Table 3. Effect of vinclozolin on the pathohistology of the testes, epididymis and bursa Fabricius in samples from laboratories 1-3. Data shown as mean scores for lesions in testes and epididymis (with number of replicates).

	Vehicle	Vinclozolin (1000 ppm)
Testes (F0 +F1)	0 (N=30)	0 (30)
Epididymis		
Duct dilation (F0)	0.8 (11)	2.1 (13)
Duct dilation (F1)	0.4 (13)	1.7 (12)
Efferent Duct Vacuolization (F0)	0.4 (8)	2.2 (11)
Efferent Duct Vacuolization (F1)	0.5 (13)	1.8 (10)