Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development

24-Jul-2008

English - Or. English

ENVIRONMENT DIRECTORATE
JOINT MEETING OF THE CHEMICALS COMMITTEE AND
THE WORKING PARTY ON CHEMICALS, PESTICIDES AND BIOTECHNOLOGY

SERIES ON TESTING AND ASSESSMENT
Number 92

REPORT OF THE VALIDATION PEER REVIEW FOR THE AMPHIBIAN METAMORPHOSIS ASSAY AND AGREEMENT OF THE WORKING GROUP OF THE NATIONAL COORDINATORS OF THE TEST GUIDELINES PROGRAMME ON THE FOLLOW-UP OF THIS REPORT

JT03249176

ENV/JM/MONO(2008)19
Unclassified

English - Or. English

OECD Environment, Health and Safety Publications

Series on Testing and Assessment

No. 92

REPORT OF THE VALIDATION PEER REVIEW FOR THE AMPHIBIAN METAMORPHOSIS ASSAY AND AGREEMENT OF THE WORKING GROUP OF THE NATIONAL COORDINATORS OF THE TEST GUIDELINES PROGRAMME ON THE FOLLOW-UP OF THIS REPORT

IOMC

INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS

A cooperative agreement among UNEP, ILO, FAO, WHO, UNIDO, UNITAR and OECD

Environment Directorate

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

Paris 2008

Also published in the Series on Testing and Assessment:

No. 1, *Guidance Document for the Development of OECD Guidelines for Testing of Chemicals (1993; reformatted 1995, revised 2006)*

No. 2, *Detailed Review Paper on Biodegradability Testing (1995)*

No. 3, *Guidance Document for Aquatic Effects Assessment (1995)*

No. 4, *Report of the OECD Workshop on Environmental Hazard/Risk Assessment (1995)*

No. 5, *Report of the SETAC/OECD Workshop on Avian Toxicity Testing (1996)*

No. 6, *Report of the Final Ring-test of the Daphnia magna Reproduction Test (1997)*

No. 7, *Guidance Document on Direct Phototransformation of Chemicals in Water (1997)*

No. 8, *Report of the OECD Workshop on Sharing Information about New Industrial Chemicals Assessment (1997)*

No. 9, *Guidance Document for the Conduct of Studies of Occupational Exposure to Pesticides during Agricultural Application (1997)*

No. 10, *Report of the OECD Workshop on Statistical Analysis of Aquatic Toxicity Data (1998)*

No. 11, *Detailed Review Paper on Aquatic Testing Methods for Pesticides and industrial Chemicals (1998)*

No. 12, *Detailed Review Document on Classification Systems for Germ Cell Mutagenicity in OECD Member Countries (1998)*

No. 13, *Detailed Review Document on Classification Systems for Sensitising Substances in OECD Member Countries 1998)*

No. 14, *Detailed Review Document on Classification Systems for Eye Irritation/Corrosion in OECD Member Countries (1998)*

No. 15, *Detailed Review Document on Classification Systems for Reproductive Toxicity in OECD Member Countries (1998)*

No. 16, *Detailed Review Document on Classification Systems for Skin Irritation/Corrosion in OECD Member Countries (1998)*

No. 17, *Environmental Exposure Assessment Strategies for Existing Industrial Chemicals in OECD Member Countries (1999)*

No. 18, *Report of the OECD Workshop on Improving the Use of Monitoring Data in the Exposure Assessment of Industrial Chemicals (2000)*

No. 19, *Guidance Document on the Recognition, Assessment and Use of Clinical Signs as Humane Endpoints for Experimental Animals used in Safety Evaluation (1999)*

No. 20, *Revised Draft Guidance Document for Neurotoxicity Testing (2004)*

No. 21, *Detailed Review Paper: Appraisal of Test Methods for Sex Hormone Disrupting Chemicals (2000)*

No. 22, *Guidance Document for the Performance of Out-door Monolith Lysimeter Studies (2000)*

No. 23, *Guidance Document on Aquatic Toxicity Testing of Difficult Substances and Mixtures (2000)*

No. 24, *Guidance Document on Acute Oral Toxicity Testing (2001)*

No. 25, *Detailed Review Document on Hazard Classification Systems for Specifics Target Organ Systemic Toxicity Repeated Exposure in OECD Member Countries (2001)*

No. 26, *Revised Analysis of Responses Received from Member Countries to the Questionnaire on Regulatory Acute Toxicity Data Needs (2001)*

No 27, *Guidance Document on the Use of the Harmonised System for the Classification of Chemicals Which are Hazardous for the Aquatic Environment (2001)*

No 28, *Guidance Document for the Conduct of Skin Absorption Studies (2004)*

No 29, *Guidance Document on Transformation/Dissolution of Metals and Metal Compounds in Aqueous Media (2001)*

No 30, *Detailed Review Document on Hazard Classification Systems for Mixtures (2001)*

No 31, *Detailed Review Paper on Non-Genotoxic Carcinogens Detection: The Performance of In-Vitro Cell Transformation Assays (2007)*

No. 32, *Guidance Notes for Analysis and Evaluation of Repeat-Dose Toxicity Studies (2000)*

No. 33, *Harmonised Integrated Classification System for Human Health and Environmental Hazards of Chemical Substances and Mixtures (2001)*

No. 34, *Guidance Document on the Development, Validation and Regulatory Acceptance of New and Updated Internationally Acceptable Test Methods in Hazard Assessment (2005)*

No. 35, *Guidance notes for analysis and evaluation of chronic toxicity and carcinogenicity studies (2002)*

No. 36, *Report of the OECD/UNEP Workshop on the use of Multimedia Models for estimating overall Environmental Persistence and long range Transport in the context of PBTS/POPS Assessment (2002)*

No. 37, *Detailed Review Document on Classification Systems for Substances Which Pose an Aspiration Hazard (2002)*

No. 38, *Detailed Background Review of the Uterotrophic Assay Summary of the Available Literature in Support of the Project of the OECD Task Force on Endocrine Disrupters Testing and Assessment (EDTA) to Standardise and Validate the Uterotrophic Assay (2003)*

No. 39, *Guidance Document on Acute Inhalation Toxicity Testing (in preparation)*

No. 40, *Detailed Review Document on Classification in OECD Member Countries of Substances and Mixtures Which Cause Respiratory Tract Irritation and Corrosion (2003)*

No. 41, *Detailed Review Document on Classification in OECD Member Countries of Substances and Mixtures which in Contact with Water Release Toxic Gases (2003)*

No. 42, *Guidance Document on Reporting Summary Information on Environmental, Occupational and Consumer Exposure (2003)*

No. 43, *Guidance Document on Mammalian Reproductive Toxicity Testing and Assessment (2008)*

No. 44, *Description of Selected Key Generic Terms Used in Chemical Hazard/Risk Assessment (2003)*

No. 45, *Guidance Document on the Use of Multimedia Models for Estimating Overall Environmental Persistence and Long-range Transport (2004)*

No. 46, *Detailed Review Paper on Amphibian Metamorphosis Assay for the Detection of Thyroid Active Substances (2004)*

No. 47, *Detailed Review Paper on Fish Screening Assays for the Detection of Endocrine Active Substances (2004)*

No. 48, *New Chemical Assessment Comparisons and Implications for Work Sharing (2004)*

No. 49, *Report from the Expert Group on (Quantitative) Structure-Activity Relationships [(Q)SARs] on the Principles for the Validation of (Q)SARs (2004)*

No. 50, *Report of the OECD/IPCS Workshop on Toxicogenomics (2005)*

No. 51, *Approaches to Exposure Assessment in OECD Member Countries: Report from the Policy Dialogue on Exposure Assessment in June 2005 (2006)*

No. 52, *Comparison of emission estimation methods used in Pollutant Release and Transfer Registers (PRTRs) and Emission Scenario Documents (ESDs): Case study of pulp and paper and textile sectors (2006)*

No. 53, *Guidance Document on Simulated Freshwater Lentic Field Tests (Outdoor Microcosms and Mesocosms) (2006)*

No. 54, *Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application (2006)*

No. 55, *Detailed Review Paper on Aquatic Arthropods in Life Cycle Toxicity Tests with an Emphasis on Developmental, Reproductive and Endocrine Disruptive Effects (2006)*

No. 56, *Guidance Document on the Breakdown of Organic Matter in Litter Bags (2006)*

No. 57, *Detailed Review Paper on Thyroid Hormone Disruption Assays (2006)*

No. 58, *Report on the Regulatory Uses and Applications in OECD Member Countries of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models in the Assessment of New and Existing Chemicals (2006)*

No. 59, *Report of the Validation of the Updated Test Guideline 407: Repeat Dose 28-Day Oral Toxicity Study in Laboratory Rats (2006)*

No. 60, *Report of the Initial Work Towards the Validation of the 21-Day Fish Screening Assay for the Detection of Endocrine Active Substances (Phase 1A) (2006)*

No. 61, *Report of the Validation of the 21-Day Fish Screening Assay for the Detection of Endocrine Active Substances (Phase 1B) (2006)*

No. 62, *Final OECD Report of the Initial Work Towards the Validation of the Rat Hershberger Assay: Phase-1, Androgenic Response to Testosterone Propionate, and Anti-Androgenic Effects of Flutamide (2006)*

No. 63, *Guidance Document on the Definition of Residue (2006)*

No. 64, *Guidance Document on Overview of Residue Chemistry Studies (2006)*

No. 65, *OECD Report of the Initial Work Towards the Validation of the Rodent Utertrophic Assay - Phase 1 (2006)*

No. 66, *OECD Report of the Validation of the Rodent Uterotrophic Bioassay: Phase 2. Testing of Potent and Weak Oestrogen Agonists by Multiple Laboratories (2006)*

No. 67, *Additional data supporting the Test Guideline on the Uterotrophic Bioassay in rodents (2007)*

No. 68, *Summary Report of the Uterotrophic Bioassay Peer Review Panel, including Agreement of the Working Group of the National Coordinators of the Test Guidelines Programme on the follow up of this report (2006)*

No. 69, *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models (2007)*

No. 70, *Report on the Preparation of GHS Implementation by the OECD Countries (2007)*

No. 71, *Guidance Document on the Uterotrophic Bioassay - Procedure to Test for Antioestrogenicity (2007)*

No. 72, *Guidance Document on Pesticide Residue Analytical Methods (2007)*

No. 73, *Report of the Validation of the Rat Hershberger Assay: Phase 3: Coded Testing of Androgen Agonists, Androgen Antagonists and Negative Reference Chemicals by*

Multiple Laboratories. Surgical Castrate Model Protocol (2007)

No. 74, *Detailed Review Paper for Avian Two-generation Toxicity Testing (2007)*

No. 75, *Guidance Document on the Honey Bee (Apis Mellifera L.) Brood test Under Semi-field Conditions (2007)*

No. 76, *Final Report of the Validation of the Amphibian Metamorphosis Assay for the Detection of Thyroid Active Substances: Phase 1 - Optimisation of the Test Protocol (2007)*

No. 77, *Final Report of the Validation of the Amphibian Metamorphosis Assay: Phase 2 - Multi-chemical Interlaboratory Study (2007)*

No. 78, *Final report of the Validation of the 21-day Fish Screening Assay for the Detection of Endocrine Active Substances. Phase 2: Testing Negative Substances (2007)*

No. 79, *Validation Report of the Full Life-cycle Test with the Harpacticoid Copepods Nitocra Spinipes and Amphiascus Tenuiremis and the Calanoid Copepod Acartia Tonsa - Phase 1 (2007)*

No. 80, *Guidance on Grouping of Chemicals (2007)*

No. 81, *Summary Report of the Validation Peer Review for the Updated Test Guideline 407, and Agreement of the Working Group of National Coordinators of the Test Guidelines Programme on the follow-up of this report (2007)*

No. 82, *Guidance Document on Amphibian Thyroid Histology (2007)*

No. 83, *Summary Report of the Peer Review Panel on the Stably Transfected Transcriptional Activation Assay for Detecting Estrogenic Activity of Chemicals, and Agreement of the Working Group of the National Coordinators of the Test Guidelines Programme on the Follow-up of this Report (2007)*

No. 84, *Report on the Workshop on the Application of the GHS Classification Criteria to HPV Chemicals, 5-6 July Bern Switzerland (2007)*

No. 85, *Report of the Validation Peer Review for the Hershberger Bioassay, and Agreement of the Working Group of the National Coordinators of the Test Guidelines Programme on the Follow-up of this Report (2007)*

No. 86, *Report of the OECD Validation of the Rodent Hershberger Bioassay: Phase 2: Testing of Androgen*

Agonists, Androgen Antagonists and a 5 α-Reductase Inhibitor in Dose Response Studies by Multiple Laboratories (2008)

No. 87, *Report of the Ring Test and Statistical Analysis of Performance of the Guidance on Transformation/Dissolution of Metals and Metal Compounds in Aqueous Media (Transformation/ Dissolution Protocol) (2008)*

No.88 *Workshop on Integrated Approaches to Testing and Assessment (2008)*

No.89 *Retrospective Performance Assessment of the Test Guideline 426 on Developmental Neurotoxicity (2008)*

No.90 *Background Review Document on the Rodent Hershberger Bioassay (2008)*

No.91 *Report of the Validation of the Amphibian Metamorphosis Assay (Phase 3) (2008)*

No.92 *Report of the Validation Peer Review for the Amphibian Metamorphosis Assay and Agreement of the Working Group of the National Coordinators of the Test Guidelines Programme on the Follow-Up of this Report (2008)*

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 30 industrialised countries in North America, Europe and the Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in ten different series: **Testing and Assessment**; **Good Laboratory Practice and Compliance Monitoring**; **Pesticides and Biocides**; **Risk Management**; **Harmonisation of Regulatory Oversight in Biotechnology**; **Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents; and the Safety of Manufactured Nanomaterials.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (http://www.oecd.org/ehs/).

This publication was developed in the IOMC context. The contents do not necessarily reflect the views or stated policies of individual IOMC Participating Organizations.

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The participating organisations are FAO, ILO, OECD, UNEP, UNIDO, UNITAR and WHO. The World Bank and UNDP are observers. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

OECD Environment Directorate,
Environment, Health and Safety Division

2 rue André-Pascal
75775 Paris Cedex 16
France

Fax: (33-1) 44 30 61 80

E-mail: ehscont@oecd.org

FOREWORD

This document contains the report of the peer review of the Amphibian metamorphosis assay performed by the United States in 2007. It is preceded by a statement from the Working Group of National Coordinators of the Test Guidelines Programme concerning the outcome of the peer review and the follow-up work.

Report of the Validation Peer Review for the Amphibian Metamorphosis Assay and Agreement of the Working Group of National Coordinators of the Test Guidelines Programme on the Follow-up of this Report

The Peer Review Report of the Amphibian Metamorphosis Assay was submitted for information to the Working Group of National Coordinators of the Test Guidelines Programme (WNT) in February 2008. Following the recommendations from the report, the WNT agreed that:

i) based on the available validation data, intra- and inter-laboratory variability should be documented in the draft Test Guideline, and performance criteria should be identified and included in the draft Test Guideline;

ii) additional guidance and details on the test conditions, exposure system, endpoint measurement, data interpretation and reference to the OECD guidance document on thyroid histopathology should be included in the draft Test Guideline to improve repeatability of the assay,

The WNT requested that the VMG-eco and its Amphibian Expert Group address technical issues identified by the Peer Review Panel or by the WNT and propose solutions to solve them, as appropriate.

Provided that the recommendations of the Peer Review Panel are addressed and considering the benefit of the Amphibian Metamorphosis Assay for the detection of substances that have thyroid agonist or antagonist activity, the WNT noted that it could provide useful information on the vertebrate thyroid system but that extrapolation from frogs to mammals is yet uncertain, and agreed to proceed to the development and finalization of the draft Test Guideline in a reasonable timeframe.

TABLE OF CONTENTS

INTRODUCTION

In 1996, Congress passed the Food Quality Protection Act (FQPA) and amendments to the Safe Drinking Water Act (SDWA) which requires 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 naturally occurring estrogen, or other such endocrine effect as the Administrator may designate."

To assist the Agency in developing a pragmatic, scientifically defensible endocrine disruptor screening and testing strategy, the Agency convened the Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC). Using EDSTAC (1998) recommendations as a starting point, EPA proposed an Endocrine Disruptor Screening Program (EDSP) consisting of a two-tier screening/testing program with in vitro and in vivo assays. Tier 1 screening assays will identify substances that have the potential to interact with the estrogen, androgen, or thyroid hormone systems using a battery of relatively short-term screening assays. The purpose of Tier 2 tests is to identify and establish a dose-response relationship for any adverse effects that might result from the interactions identified through the Tier 1 assays. The Tier 2 tests are multi-generational assays that will provide the Agency with more definitive testing data.

One of the test systems recommended by the EDSTAC was the Amphibian Metamorphosis Assay (AMA). The AMA consists of multiple endpoints; principally, developmental stage, hind limb length, body length (whole body and snout-vent), histology of the thyroid glands, mortality and morbidity. It is intended to empirically identify substances which may interfere with the normal function of the hypothalamic-pituitary-thyroid (HPT) axis. It represents a generalized vertebrate model to the extent that it is based on the conserved structure and functions of thyroid systems. It is an important assay in the EDSP screening battery because amphibian metamorphosis provides a well-studied, thyroid-dependent process which responds to substances active within the HPT axis, and it is the only assay in the battery that assesses thyroid activity in an animal undergoing morphological change.

Although peer review of the AMA will be done on an individual basis (i.e., its strengths and limitations evaluated as a stand alone assay), this assay, along with a number of other in vitro and in vivo assays, will likely constitute a battery of complementary screening assays. A weight-of–evidence approach will also be used among assays within the Tier-1 battery to determine whether a chemical substance has the potential to interact with the endocrine system and whether Tier-2 testing is necessary. Peer review of the EPA's recommendations for the Tier-1 battery will be performed at a later date by the FIFRA Scientific Advisory Panel (SAP).

The purpose of this peer review was to review and comment on the amphibian metamorphosis assay for use within the EDSP to detect chemicals which may interfere with the normal function of the hypothalamic-pituitary-thyroid (HPT) axis. The primary product peer reviewed for this assay was an Integrated Summary Report (ISR) that summarized and synthesized the information compiled from the validation process (i.e., detailed review papers, pre-validation studies, and inter-lab validation studies, with a major focus on inter-laboratory validation results). The ISR was prepared by EPA to facilitate the review of the assay; however, the peer review was of the validity of the assay itself and not specifically the ISR.

The remainder of this report is comprised of the unedited written comments submitted to ERG by the peer reviewers in response to the peer review charge (see Appendix A). Section 2.0 presents

peer review comments organized by charge question, and Section 3.0 presents peer review comments organized by peer review expert. The Integrated Summary Report is presented in Appendix B and additional supporting materials are included in Appendix C.

The final peer review record for the amphibian metamorphosis rat assay will include this peer review report consisting of the peer review comments, as well as documentation indicating how peer review comments were addressed by EPA, and the final EPA work product.

Peer Review Logistics

ERG initiated the peer review for the amphibian metamorphosis assay on October 24, 2007. ERG held a pre-briefing conference call on November 16, 2007 to provide the peer reviewers with an opportunity to ask questions or receive clarification on the review materials or charge and to review the deliverable deadlines. Peer review comments were due to ERG on or before December 5, 2007.

Peer Review Experts

ERG researched potential reviewers through its proprietary consultant database; via Internet searches as needed; and by reviewing past files for related peer reviews or other tasks to identify potential candidates. ERG also considered several experts suggested by EPA. ERG contacted candidates to ascertain their qualifications, availability and interest in performing the work, and their conflict-of-interest (COI) status. ERG reviewed selected resumes, conflict-of-interest forms, and availability information to select a panel of experts that were qualified to conduct the review. ERG submitted a list of candidate reviewers to EPA to either (1) confirm that the candidates identified met the selection criteria (i.e., specific expertise required to conduct the assay) and that there were no COI concerns, or (2) provide comments back to ERG on any concerns regarding COI or reviewer expertise. If the latter, ERG considered EPA's concerns and as appropriate proposed substitute candidate(s). ERG then selected the five individuals who ERG determined to be the most qualified and available reviewers to conduct the peer review.

A list of the peer reviewers and a brief description of their qualifications is provided below.

- ***David Crews, Ph.D.,*** is the Ashbel Smith Professor of Zoology and Psychology at the University of Texas at Austin. His research primarily concerns sex determination and sexual differentiation in nonmammalian vertebrates, with a special focus on the role of the physical and biotic environment on these fundamental processes. He has published influential articles on the issue of threshold and mixtures as they relate to endocrine disrupting compounds. Dr. Crews has worked with a wide variety of organisms, from fruit flies to mammals, but focuses on reptiles. He has published over 350 research articles, book chapters, and essays in the areas of reproductive biology, neuroscience and endocrinology and edited four books. His published papers have appeared in *Science*, *Nature*, and the *Proceedings of the National Academy of Sciences*. He has received a number of honors, including a Research Scientist Award (1977-1997) and a MERIT Award from the NIH. He is fellow of the American Academy of Arts and Sciences and other societies.

- ***J. David Furlow, Ph.D.,*** is an Associate Professor at the University of California, Davis, in the Section of Neurobiology, Physiology, and Behavior, College of Biological Sciences, where he has served on the faculty since 1998. Dr. Furlow received his bachelor's degree in Biochemistry from the Pennsylvania State University and his Ph.D. in Biochemistry at the University of Wisconsin. At Wisconsin, Dr. Furlow did his thesis work on estrogen receptor structure and function with Dr.

Jack Gorski. His post-doctoral training was done at the Carnegie Institution of Washington Department of Embryology, in the laboratory of Dr. Donald Brown. It was at the Carnegie Institution where he began working on the problem of thyroid hormone control of metamorphosis in Xenopus laevis, research that is ongoing in his current laboratory at Davis. The Furlow lab investigates the control of gene expression by the thyroid hormone receptors during metamorphosis, with additional related interests in the impact of environmental chemicals in modulating thyroid hormone receptor activity and the development of synthetic thyromimetic compounds. In addition to research on thyroid hormone action, Dr. Furlow has an active collaboration with Dr. Sue Bodine at Davis on reciprocal control of skeletal muscle atrophy by corticosteroids and IGF-1 in rodent models. Dr. Furlow has served as a faculty member and director of the Physiology summer course at the Marine Biological Laboratory in Woods Hole, MA and most recently as a faculty member at a gene expression course run by the European Molecular Biology Laboratory in Heidelberg, Germany. He has authored 26 articles on estrogen and thyroid hormone function, and has served on several ad hoc grant review committees for both the National Institutes of Health and the National Science Foundation. Research in Dr. Furlow's laboratory is currently funded by grants from the National Institutes of Health, the Muscular Dystrophy Association, and the Netherlands Organization for Scientific Research.

- ***Catherine Propper, Ph.D.,*** is a Professor in the Department of Biological Sciences at Northern Arizona University. She has 25 years of experience investigating the interactions between the environment and endocrine physiology and behavior using amphibians as model systems. Her current research focuses on the impacts of environmental endocrine disruptors on development, reproduction and behavior using *Xenopus laevis* and *Xenopus tropicalis* as models. Her work investigates the impact of exposure to single compounds and the complex chemical mixes found in wastewater effluent. Dr. Propper has served on National Science Foundation and Environmental Protection Agency grant review panels. She is a Faculty Affiliate of the Arizona Water Institute (AWI), a consortium of Arizona Universities, State Government and private stakeholders working together to solve the complex issues associated with water availability in the arid Western U.S., and she serves on Northern Arizona University's Executive Committee for the AWI. She is also a Faculty Associate with the Merriam Powell Center for Environmental Research and is a Core Principle Investigator with the Center for Discovery Research both at Northern Arizona University. She has developed graduate courses in Endocrine Disruption. Her federal and state funded research activities include investigating the mechanism of octylphenol disruption of gonadal development, impacts of endocrine disrupting compounds on behavior, and development of rapid assay assessment tools for chemicals in wastewater. She publishes her work in journals such as *Environmental Health Perspectives, General and Comparative Endocrinology, Hormones and Behavior, Journal of Environmental Biology, and Journal of Experimental Biology*.

- ***Hannes van Wyk, PhD.,*** is a professor in Zoology at the University of Stellenbosch in South Africa. He has been associated with the University of Stellenbosch since 1988, teaching general biology, histology, animal physiology and aspects of ecotoxicology. He has supervised several postgraduate students. Initially his research focused in the field of Herpetology, specifically on comparative reproduction and endocrinology of local reptilian species and as well as androgen controlled epidermal glands in lizards and amphibians species. In 1998 he was part of the initiative to develop an EDC programme in South Africa and subsequently was the principle investigator for several projects related to EDC biomarker development and funded by the SA Water Research Commission (WRC) and National Research Foundation (NRF), specifically using *Xenopus laevis* as bio-indicator species. Initial research focused on the natural reproductive cycles in local *Xenopus laevis* populations and the development of bioassays for estrogen-response proteins (including vitellogenin). An *in vitro X. laevis* liver slice bioassay for estrogenicity screening using monoclonal anti-Vtg antibodies as well as an universal vertebrate VTG ELISA

was developed and validated. Male *X. laevis* nuptial skin glands were validated as biomarkers for androgenic/anti-androgenic activity. During 2006, WRC funded a project to establish a *Xenopus* metamorphosis assay locally to screen for thyroid disrupting activity in water sources. Prof van Wyk has used the XEMA protocol on two occasions to screen environmental water samples for thyroid activity. In order for local ecotoxicologists to use XEMA, Prof van Wyk is currently busy to compile an Atlas for *Xenopus laevis* development including changes in thyroid histology during metamorphosis. In addition to the basic XEMA protocol, his group established QPCR methodology to study the expression of female related genes during juvenile development in tilapia fish and to study relative gene expression of thyroid related genes (TRβmRNA) in *X. laevis*. Gene expression studies related to androgen- and thyroid hormone receptors in tilapia fish is also underway. In addition, research related to the seasonal reproductive biology and sex determination in other fresh water species, a local freshwater turtle (*Pelomedusa subrufa*) and the Nile crocodile is ongoing. Contract research was published in peer reviewed WRC reports or submitted to local and international journals.

- ***Richard Wassersug, PhD.,*** is a Full Professor in the Department of Anatomy and Neurobiology at Dalhousie University in Halifax, Nova Scotia as well as a Research Associate at both the Field Museum of Natural History in Chicago and the Smithsonian Institution, Washington DC. The majority of his scientific career has been spent studying the functional morphology and behavior of anuran larvae; he has published approximately 100 peer-reviewed papers on these animals. Approximately a quarter of his peer-reviewed papers are on the behavior, morphology, physiology, and development of *X. laevis*. He maintained a laboratory colony of *Xenopus laevis* for 30 years and has published on standard operating procedures for the care and maintenance of this species.

PEER REVIEW COMMENTS ORGANIZED BY CHARGE QUESTION

Peer review comments received for the amphibian metamorphosis assay are presented in the sub-sections below and are organized by charge question (see Appendix A). Peer review comments are presented in full, unedited text as received from each reviewer.

Overall General Comments

General comments provided by several reviewers are summarized below.

David Crews: There is much to praise about this report, in particular the careful thought and precision of the experimental protocol in all three phases of the process. However, it is the opinion of this reviewer that the conclusions regarding inter-laboratory variability are not warranted and that it fails as a method for accomplishing the stated goal of the assay to be part of the Endocrine Disruptor Screening Program (EDSP). This assessment is based on the fact that endocrine disrupting compounds are rarely (if ever) found in nature as the sole contaminant, that such mixtures interact in a manner that must be tested before the interactions can be discarded as factors, and that endocrine disrupting compounds/chemicals (EDCs) act on integrated endocrine systems during development that have consequences beyond the life history of the individual organism. As a traditional environmental toxicology exercise, the assay is a first step, but still ignores the issue of low dosages and the need for other endocrine endpoints.

Catherine Propper: This assay was developed to determine whether compounds to be testing for Tier 1 Level analysis in the EPA's Endocrine Disruptor Screening Program disrupt thyroid hormone function. Amphibians are an outstanding model for investigations of thyroid hormone function because the process of metamorphosis is strongly regulated by first the expression of the thyroid hormone receptor and then later the secretion of thyroid hormones from the thyroid gland. Therefore, compounds that mimic thyroid hormone activity may increase the rate of metamorphosis, and those that antagonize thyroid hormone activity or function can decrease the rate of metamorphosis. Clear morphological and developmental endpoints are readily evaluated to determine the impact of exposure. Therefore this assay is readily transferable doable across laboratories. The utility of the assay also makes it functional for non-contracted investigators to study chemicals and complex mixes that may not be under the purview of the Endocrine Disruptor Screening Program.

The validation of the Amphibian Metamorphosis Assay (AMA) involved three phases of validation. The first phase investigated how differences in exposure timing could impact outcomes and whether there was significant interlaboratory variation in outcomes. A multichemical study was also undertaken by the USEPA using both exposure timing scenarios. The second phase involved used the information derived from Phase I to formulate a standard operating document. This assay was then used compare exposure outcomes to several compounds with predicted thyroid or antithyroid activity across several labs. The third phase of the study evaluated a compound with strong endocrine activity (estradiol), but predicted not have direct thyroid hormone activity (please see comments below), and one with weak activity as evidenced in some literature. The validation studies demonstrate overall the utility of this assay for evaluating thyroid disrupting activity of the compounds tested. My comments below 1) address needed details in the final AMA Test Methodology, and 2) describe the limits of the assay as it was performed in the validation steps.

In reviewing the materials for the Amphibian Metamorphic Assay, I have followed the review Guidelines provided by the EPA. Some of my comments are general and not referenced to the page number on the Integrated Summary Report (ISR) and three Test Method Documents, and some are specifically referenced. Under section 8, I summarize my main criticisms of the AMA based on what issues that I evaluated under specific sections.

Richard Wassersug: (*General comments are extracted from an e-mail sent to ERG on December 4, 2007*)

"…I have focused largely on the care and welfare of the tadpoles. That is more my area of expertise than either the toxicological implications of the chemical agendas used in the various tests for developing the assay, or the histopathology of the thyroid gland itself. Those two areas seem solid and strong. My major concerns center around better ways to standardize the care of the tadpoles such that their level of stress can be minimized or at least consistent across labs.

My review follows the order of the eight questions laid out in the "charge to reviewers." I critique the Summary document on the phase I, II, and III trials, as well as the methodological document on how one executes the AMA. Since my review is rather long, I have placed in bold the key sentences expressing my major concerns. Please let me know as soon as possible if you need any additional information from me to justify my assessment.

I favor the EPA accepting the AMA as one of its core assays for endocrine disruptors. But at the same time I feel that the issues raised in my review need to be addressed. If appropriate, I am willing to work with officials at the EPA to explore ways to make the methodology for the execution of the AMA more rigorous and reliable.

Thank you for giving me an opportunity to participate in this review process…."

Comment on the Clarity of the Stated Purpose of the Assay.

David Crews: The documents provided document the rationale for an amphibian metamorphosis assay (**AMA**) as a high throughput *in vivo* assay for thyroid disrupting chemicals. A series of tests designed to validate this method are described using the tadpole *Xenopus laevis*.

The document "Integrated Summary Report – Amphibian Metamorphosis Assay" (File Name: Ama_isr) presents a protocol designed such that an aquatic toxicology laboratory would be able to conduct studies of chemicals for their effects on the developing thyroid system of this animal model system.

Specifically, tadpoles reared under standardized conditions will be treated during a discrete period of development beginning at Stage 51 will be exposed for 21 days to one of several concentrations of the test chemical; another group will be exposed to a water control. Within each chemical treatment there will be four replicates. At each of three time points (d0, d7 and d21 or treatment) the endpoints measured will include developmental stage, wet weight, snout-to-vent length (SVL), whole body length (WBL), hind limb length (HLL), and thyroid histology. The latter two measurements will utilized dissecting (limb length) or light microscopic measurements with computer-assisted image-digitizing software measurements. Finally, tadpoles will be observed daily for mortality and malformations.

A flow-through method for delivering the chemicals at the various concentrations will be used with measurements being taken at periodic intervals (weekly) to evaluate and validate the composition of the water. It appears that the preferred system will require that each set of 4 replicate tanks (= test vessel) will receive a given concentration using a diluter system. It is commendable that in this method the test tank

will not serve as a feed to other tanks. The alternative method, static-renewal, is not described and so cannot be evaluated by this reviewer.

Each replicate tank will be a 4 litre glass aquarium with 20 larvae initially. Light, temperature, pH, DO, and feeding will be standardized, with the tanks randomly situated to allow for possible differences due in placement.

Adult male and female South African clawed frog *Xenopus laevis* will be injected in human chorionic gonadotropin (hCG) to induce breeding. The source of the adult animals (pg. 5), and the "best spawns" (pg. 5) are a concern (see 2. below). The larvae will be raised in constant densities, being fed twice daily during the week and once daily on weekends and holidays.

Three test concentrations will be utilized. The highest, the maximum test concentration (MTC), is defined as the highest test concentration of the chemical that results in less than 10% acute mortality. This is a concern (see below). The lower concentrations to be tested would be calculated as a dose separation of 0.33-0.5 (max-min).

Test animals will be selected on the basis of normal body morphology and using the hind limb morphology staging criteria of Nieuwkoop and Faber (pg. 10). For a d0 measure, approximately 20 individuals will be measured for WBL. It is not clear if these 20 individuals will be reintroduced into the test population for distribution into the tanks, or whether they will be used to obtain the other stated measurements (see above).

A sample of 5 tadpoles will be taken from each tank on d7, for a total sample size of 20 tadpoles for each treatment/dose, and a detailed selection procedure is outlined for obtaining a similarly sized sample on d21.

The histological measurements are described well as are the statistics to be applied. Procedures for Data Reporting are similarly clear, but should be made mandatory, rather than recommended. It is not clear what is meant by "gross deviations from the test method" and so cannot be evaluated. This should be rigorously defined.

The document "Guidance Document on Amphibian Thyroid Histology Part 1: Technical guidance for morphologic sampling and histological Preparation" (File Name: AMA_Test_Method_Appendix_1), is overall outstanding in its instructional clarity regarding the handling and euthanasia of tadpoles, biometry, and preparation of the samples for analysis. As one who has over 35 years of experience in all aspects of histology, especially paraffin processing, I cannot think of anything that has not been anticipated. There is one important factor that is omitted, however, is consideration of asymmetrical limbs (see below).

The document "Guidance Document on Amphibian Thyroid Histology Part 2: Approach to reading studies, diagnostic criteria, severity grading, and atlas" (File Name: AMA_Test_Method_Appendix_2), is also outstanding in its instructional clarity, breadth and depth. I am impressed by the careful attention to avoiding errors in assessment in addition to the more standard diagnostic criteria for grading slides. The section images themselves are outstanding in both magnifications and the histologist who prepared them is to be congratulated. However, there is a serious flaw in Section IB. Approach to reading studies (see below) regarding the scoring of the slides.

J. David Furlow: The purpose of the assay is to screen for environmental compounds that affect the hypothalamus-pituitary-thyroid axis, using an intact animal model. Overall the document is clear, and the amount of work setting up and evaluating the system is very impressive. Indeed, a standardized method for raising *Xenopus laevis* through metamorphosis for this level of analysis has been surprisingly lacking. The advantages of the system are clear: the system has dramatic, easily measured external morphological changes to a hormone that is identical in structure to its human counterpart. Furthermore, the assay is conducted in a developing animal as opposed to the other battery of whole animal assays the EPA is considering that are conducted in pubertal or adult rats (pubertal male and female rat assay; ovariectomized female rat assay).

The one statement I would add to the stated purpose section is that the assay can also detect disruption of thyroid hormone signaling at the target cell i.e. the presence of thyroid hormone receptor agonists or antagonists (especially since the recommended starting stage is 51 prior to the precence of detectable circulating TH). As stated, the implication is that the assay will only detect disruption of the pathways controlling thyroid hormone synthesis.

Catherine Propper: ISR Pages 12-13: The overview and justification within the ISR is a brief review describing why the amphibian system provides a strong assay for investigation of the potential for anthropogenic compounds to impact thyroid related function. One addition that would be useful for this summery is a stronger overview of the timing of expression of thyroid hormone receptors during development compared with the release of thyroid hormone from the thyroid gland during amphibian metamorphosis. Such an explanation helps in the understanding of the set up of the two assay regimes that were tested in the Phase I validation trials. Second, a brief overview of the receptors repressor verses activator activities might be useful ultimately for interpretation of outcomes, and because the receptors are expressed prior to increases in TH secretion. This information is critical to the understanding of the timing of the assay because the expression of TR during the earlier stages of the assay period (51-53) may lead to repression of TH sensitive genes and allow instead for growth of the tadpoles during this period, but if an environmental mimic is present, it could shift the activity of the receptor and accelerate metamorphosis. Buchholz *et al* (2006) is a useful review.

Hannes van Wyk: The Introduction and stated purpose of a Tier 1 assay was clear. Personally I think the general explanation of the purpose of a Tier 1 assay is extremely important. I don't think it is always appreciated what the actual purpose of a Tier 1 assay is. In the Introduction and background to the stated purpose of the assay the progression of assay development, validation and evaluation are important components to the reader. In order to underline the role/place of a Tier 1 screening assay in the larger picture of assessing EDC activity I would like to see a diagramme showing the contribution of Tier 1 screens. The criteria set by EDSTAC for Tier 1 screens were presented. With this statement "*It is important to recognize that the AMA is not intended to quantify or to confirm endocrine disruption, or to provide a quantitative assessment of risk, but only to provide suggestive evidence that thyroid regulated processes may be sufficiently perturbed to warrant more definitive testing*" the purpose of the AMA is placed within the framework of a Tier 1 screening programme underlining the purpose of such an assay and sets the scene to understand the development and validation of a Tier 1 assay.

Richard Wassersug: The purpose of the Amphibian Metamorphosis Assay (AMA) is clearly stated in the EPA documents.

Comment on the Clarity, Comprehensiveness and Consistency of the Data Interpretation with the Stated Purpose of the Assay

David Crews: As instructed (pg. 12) in the document "FINAL REPORT OF THE VALIDATION OF THE AMPHIBIAN METAMORPHOSIS ASSAY FOR THE DETECTION OF THYROID ACTIVE SUBSTANCES: PHASE 1: OPTIMISATION OF THE TEST PROTOCOL" (File Name: OECD_Phase_1_Report.pdf), this report and that following "FINAL REPORT OF THE VALIDATION OF THE AMPHIBIAN METAMORPHOSIS ASSAY: PHASE 2: MULTI-CHEMICAL INTERLABORATORY STUDY" (File Name: OECD_Phase_2_Report.pdf) will be considered together. However, I will focus on the data from the stage 51, 21 d treatment group for Phase 1 since that is the developmental stage for the initiation of treatment in the AMA protocol.

PHASE 1 REPORT

Summary *i)* It is stated that the origin of the effort to develop and optimize an AMA "originated at a meeting of the Amphibian Expert Group, an advisory group to the Validation Management Group, in June 2003 at a meeting hosted by the US Environmental Protection Agency in Duluth, MN, USA." (**pg. 18**) There is no reference to another EPA-sponsored workshop (DeVito et al., 1999). This is unfortunate because a specific observation/caution was made (Thyroid function affects reproductive development and function). Further, a specific recommendation appears to have been ignored in the present effort. Namely, "A number of assays or test systems can be used to detect chemicals that produce hypothyroidism. However, most of these assays or test systems are time consuming and not necessarily specific for hypothyroidism. In addition, pronounced decreases in serum T4 concentrations are required to detect the behavioral or morphologic changes. Alterations in serum THs can be detected at lower dose levels than those required to detect the behavioral and morphologic changes in these systems. Because of the greater sensitivity and simplicity, determination of serum TH concentrations is recommended instead of these developmental assays. It should be remembered that using adult, pubescent, or prepubescent animals may be qualitatively predictive of fetal response, whereas it may not be quantitatively predictive of dose or response in fetal tissue." (pg. 412 of DeVioto et al., 1999).

Summary *iv* and *vi)* In the first phase three participating laboratories each used "their specific methods to test the anti-thyroid compound, 6-propylthiouracil (PTU), and the receptor agonist, T4, at comparable exposure concentrations." In the second phase identical methods were used by six participating laboratories with a total of 14 experimeintal studies with the replication of T4, and two new chemicals, specifically sodium percholorate (Na-PER), a thyroid hormone synthesis inhibitor, and iopanoic acid (NIS), a deiodinase inhibitor.

Statistical Analysis (pg. 23, Phase 1).

Gene expression (item 14). It is not appropriate to simply presume that the gene expression data followed a log-normal distribution. It should first be tested for heterogeneity of variance and then, if appropriate, the transform done. Further, there description of the methodology for the semi-quantitative RT-PCR ("densitometric analysis of scanned agarose gels are shown. Results were expressed relative to the control Group") is not adequate. Show me the protocol and the original data so that I can determine the validity of the method.

Analytic Chemistry Results Standard Deviations (Tables 3 and 6). A replicate is defined on Pg. 22 and described as "20 tadpoles were used per replicate tank in the GER and JPN laboratories; the US laboratory used 25 tadpoles per replicate tank in the PTU studies." Considering only the Stage 51 sutdy,

the variability in PTU concentrations in the US laboratory is commendable, but that of the JPN laboratory is of concern. This is amplified in the lack of a 0.00 concentration in the JPN measurements, raising the question of whether their control water actually has compounds that cross react in the measurement system. A similar problem exists for the T4 concentrations (Tables 4 vs. 7).

Item 21. Comparison of Control Data (pg. 25). This is a misrepresentation as there is no data provided by the GER laboratory, and that of the JPN laboratory is questionable.

Table 8. Consideration of the median is misleading in that the tadpoles from the GER laboratory have a bimodal distribution of development for the PTU, and develop slower under the T4 regimen.

Table 11. It is extremely odd that in the JPN laboratory control tadpoles from the two treatment groups varied substantially (there is no overlap by one STD).

Item 27 and Table 12 (pgs. 28 and 29). "The significant difference at 5 mg/L after 14 days of exposure in JPN study seems to be an anomalous result and driven by one of the two replicates which does not fit the pattern of the other tests." Considering the above comment under Item 21, this may not be so anomalous and should not be disregarded. It is this reviewer's opinion that the absence of analytic chemistry of the GER lab, and the questionable quality of the analytic chemistry of the JPN lab, there is really no points of comparison from the null condition.

Table 13 vs. Table 15 Comparison. These tables present data from two laboratories (GER and JPN, respectively) for the same treatment conditions. However, if we compare the information for hind limb length for the GER lab, we see that the difference in Pool means between d7 and d21 values are:

	Control	**2.5 mg/L**	**5.0 mg/L**	**10 mg/L**	**20 mg/L**
GER	8.9	7.6	6.6	6.8	6.4
JPN	10.6	11.4	8.7	10.2	3.2

It does not take a scientist to come to the conclusion that the data produced by the two laboratories are not comparable.

Figure 2. The substantial SEMs in the d21 TSHb and BTEB are troublesome.

Tables 27, 29 and 31 Comparison. The only valid measure of inter-laboratory concordance is that of body weight. Comparing the difference between the control and the 2.0 mg/L T4 average values for the GER, JPN, and US sites are: 274, 205, 402, respectively, this and inspection of the trends within each lab, I conclude that they cannot be compared.

Conclusion: Phase 1 data is not valid in terms of inter-laboratory comparison. While **"**these studies resulted in remarkably similar outcomes among the different laboratories, despite minor methodological differences", the results from each laboratory cannot be combined one with the other, severely limiting any attempts at meta-analysis.

PHASE 2 REPORT

Summary (pg. 19). The purpose of the " Phase 2 of the validation study aimed at an inter-laboratory multi-chemical testing with an harmonised protocol." Specifically, tadpoles reared under standardized conditions were treated during a discrete period of development beginning at Stage 51 for 21 days to one of concentrations of the test chemical; another group will be exposed to a water control. Within each chemical treatment there will be four replicates. At each of three time points (d0, d7 and d21 or treatment) the endpoints measured will include developmental stage, body mass as wet weight, SVL, WBL, and HLL as

well thyroid histology. Six international laboratories performed a total of 14 studies using Na-PER (n=4) and IOP (n=4).

Identity of Laboratories. This information is not provided and this is very regrettable. It is vital to know if any given laboratory can reproduce its data for certain controls, in this instance the no chemical group and the T4 group. If one or more of the three laboratories in the Phase 1 study participated in the Phase 2, this would enable evaluation of QC/QA. This point is further evidenced in the finding (Tables 8 and 9) that "The intra-laboratory comparison of tadpole growth parameters showed highly reproducible results in lab 1, lab 2 and lab 5 and less reproducible results in lab 3." (pg. 34)

Growth in the Control Group (pp. 33-40) While reassuring, the finding that tadpole growth within the control groups within a particular laboratory are reproducible, this is not at all satisfactory if the aim is to be able to compare across laboratories. Table 9 in particular would convince any reviewer for a reputable scientific journal to recommend rejection.

Effects of Na-PER and IOP on Developmental Endpoints. If it is not possible to compare the laboratories in terms of the control group, then there is no point in attempting to make sense of the inter-laboratory variation in the experimental groups. In this regard, lab 1 has a reasonable dose-response curve for Na-PER at d21 for WBL, SVL, and mass (PER (Tables 12-14).

Items 55, 64 and 77. The presentation of results for histopathology of two laboratories that are not comparable is misleading at best. What kind of conclusion can be drawn from this data?

Effects of T4 on Developmental Endpoints. The comparison of Tables 23-26 suggest that tadpoles in laboratories 1 and 2 showed limb growth but little or no change in mass, whereas animals in , the animals limbs responded but not mass, whereas for laboratories 3 and 4 the opposite pattern existed.

Conclusion. Phase 2 was conducted in an exemplary fashion in terms of standardizing protocols. The conclusion that "these studies resulted in remarkably similar outcomes among the different laboratories, despite minor methodological differences" is certainly true within each laboratory. However, the results from each laboratory cannot be combined one with the other, severely limiting any attempts at meta-analysis. Thus, the most important opportunity this Phase allowed, namely the comparison across laboratories, is an unqualified negative. Finally, it is vital that any laboratories that participated in both Phases 1 and 2 be compared for control group measurements.

PHASE 3 REPORT

Summary. In the Phase 3 study additional compounds were recommended for study, benzophenone-2 (BP-2), 17β-estradiol (E2), potassium iodide (KI) and p,p'-DDE (DDE), but experiments were only conducted on BP-2 and E2. However, the concept of including both positive and negative controls in Phase 3 is excellent.

Control group. Inspection of Table 2 and the statement on pg. 16 "there was no solvent control" suggesting there was no control group in this study. If this indeed were the case, then no conclusions can be drawn about the relationship between BP-2 and E2. This clearly is an omission, but an important one in Table 2..

Statistical Analysis. If there is no control group, what is the basis of the statement on pg 16 "Dunn's test was used for pairwise comparisons of treatment group medians to the control median" and on pg. 17 "pairwise comparisons of treatment group means to the control mean were performed using Dunnett's/Tamhane-Dunnett's test."

Growth in the Control Group. As stated (pg. 22) "Control tadpoles used in the two independent experiments in lab 1 showed similar growth rates indicating low intra-laboratory variability. In comparison to lab 1, control tadpoles used in the experiment performed in lab 2 were greater in size, as judged from WBL and SVL measurements on day 7, and had increased body weights." However, Tables 3-5 do not contain data clearly labeled as the 0.0 or DWC group. While this can be understood for Table 3 (as it is d0), the legends for Tables 4 and 5 as well as for Figures 2 and 3 caused this reviewer considerable time and effort before understanding that they were misstated.

Sex Determination. It is not clear how sex assignment was determined. What were the criteria used in the "gross morphological assessment" (pg. 26)?

Table 7 (pg. 28). It should be noted that the effect of the 2.0 and 10 mg/L E2 is most likely due to the reduction in the variance, which almost certainly is due to the larger sample size.

Discussion. The interaction of the thyroid system is presented as unidirectional and cause-effect (pg. 51), that is how gonadal steroid hormones affect the pituitary-thyroid axis or with TH action. This is misleading. First, it does not consider how the thyroid might affect the developing gonad. Second, the emphasis should be on the interaction of two axes during development, namely the hypothalamo-pituitary-gonad and the hypothalamo-pituitary-thyroid axes.

It is understood that interpretations of the literature are prone to the biases of the reviewer. This reviewer disagrees with the statement of the authors of Phase 3 that "interference of gonadal steroids with the thyroid system occurs most likely at the hypothalamic-pituitary level" (pg. 52) and present additional evidence in the section 5. Limitations. f.

Conclusion. Overall a good study. Consideration however should be given to the issues identified above.

J. David Furlow: The endpoints of the assay are stated as the following: mortality, hindlimb length, whole body and snout vent length, developmental stage (although this is primarily based on fore- and hind-limb size and morphology during the stages analyzed), body weight, and thyroid gland histology.

However, I am concerned that the stage 51, 21 day assay is not sufficiently comprehensive or sensitive to detect interference with the HPT axis. Control animals (both in the Phase I and Phase II trials) only usually progress to stage 58 or 59. This precludes any consideration of compounds that affect tail resorption that demands attainment of the highest levels of T3 in the tissue to respond. As an example, overexpression of prolactin does not inhibit any observable aspect of progression through metamorphosis except for resorption of the connective tissue of the tail (Huang H, Brown DD. Prolactin is not a juvenile hormone in Xenopus laevis metamorphosis. Proc Natl Acad Sci U S A. 2000 97(1):195-9.). Perhaps even more relevant, transgenic overexpression of the Type III deiodinase that degrades T4 and T3 arrests animals between stages 60 and 61 with the most obvious effect on gill and tail resorption (Huang H, Marsh-Armstrong N, Brown DD.Metamorphosis is inhibited in transgenic Xenopus laevis tadpoles that overexpress type III deiodinase. Proc Natl Acad Sci U S A. 1999 96(3):962-7.). Limb growth was not affected in this experiment.

The conclusion that the assay is sufficiently reproducible between laboratories will be addressed under item 7.

Catherine Propper:
a. The endpoints are clear, not difficult to monitor, and appear to provide fairly consistent results across the validations. Some specific comments are below, however, regarding the interpretation of the data.

b. In the ISR, three possible outcomes are delineated on Page 22 Section 3.6 under Data Interpretation: "Thyroid Active, Thyroid Inactive, and toxic." The problem with this wording it that "thyroid active" does not distinguish between whether a compound is acting like thyroid hormone or inhibiting thyroid hormone activity. A possible change would be to have 4 categories (breaking up the first category in the original document into two categories that represent the different forms of "thyroid activity"). One possible suggestion would be "Thyroid mimic" and "Anti-thyroid Activity."

c. Sensitivity section page 49 ISR: The section in the ISR that tries to summarize which assay is most sensitive (14 verses 21 day) is not that clear. Although after several passes through the table, I was able to come to the same conclusion as the ISR, a brief summary table for sensitivity would be useful.

d. After Phase I, the decision was made to use flow-through systems not only for the other phases of the study, but also in the final AMA Test Method. However, no justification is provided for deciding to use the flow-through system. In other words, no statistical comparison was made to determine that this provides the better means of getting a more sensitive result (see further comments below).

e. A much stronger guideline for data interpretation within the AMA Test Method Documents is necessary. This issue was brought out when evaluating the Phase III estradiol results. In this trial, there lack of consistency in interpreting the estradiol exposure results when compared with the interpretation from phase I and II trial results. For example, the Phase II summary Table 6.1 in the ISR, Table X says there is no developmental effect, and then the report goes on to state that there is a significant reduction in the number of tadpoles reaching stage 60 in the estradiol groups. Is there an effect or not? This result suggests that investigators also should determine the number of animals not reaching a specific stage when conducting the AMA methodology. What was the statistical evaluation on the developmental stage across all groups? In the phase III study, clearly, more animals reached stage 60 in the controls than in the higher E2 doses (this finding is supported in a couple of papers in the literature in *Xenopus (eg.Gray & Janssens 1990))*, suggesting minimally, that E2 is interfering with thyroid hormone activity although the mechanism is not well understood. Also inconsistent is the fact that also found was a decrease in hind limb length which in the Phase III trial is considered to be general toxicity, but in the other toxicity measures are considered to be negative for toxicity. For example, these same findings in phase I and II would have lead to an interpretation of thyroid hormone antagonistic activity of E2. Such interpretation suggests that the data were evaluated based on the expected result for estradiol not being a "thyroid active" compound rather than on the outcome of the data. Last, there is a strong literature on the interaction between thyroid hormone and estradiol in mammals (Pfaff *et al.* 2000) and the receptors interact in ways that are complex (Vasudevan *et al.* 2002). This information suggests that, in fact, the Phase III trial demonstrated that estradiol may have thyroid disrupting activity. The Phase III results are very consistent with that literature and should be reinterpreted both to be consistent with the phase I and II studies and in light of this literature. This issue of data interpretation comes up again in the Phase III BP-2 studies which also suggest that the effects of BP-2 on thyroid hormone function could be confounded by its direct actions and indirect actions because it also may act as an estrogen. And last, to further support this inconsistency in interpretation, in the IOP experiment of Phase III, lab 2 had the exact same findings (including decreased developmental stage and no thyroid histopathology impact) and these data are interpreted to be "thyroid active." *In summary, this phase trial demonstrated that data interpretation across the validation studies needs to be consistent, and guidelines need to be carefully developed to facilitate this interpretation.* In fact, in the AMA Test Method, there is no section on data interpretation, and in the overall ISR, there are no clear guidelines for how many parameters need to be significantly different from controls before a compound is to be interpreted as thyroid disrupting. Such guidelines are essential and should be provided clearly in the final AMA Test Method Protocol, along with appropriate summary tables.

Hannes van Wyk: Data interpretation in some cases was difficult. It is acknowledged that in terms of the size related endpoints non-thyroidal effects may have been operational. Throughout the authors tried to do

comprehensive interpretations and were mostly consistent focusing on the fact the AMA is a screen for thyroid interaction. They must be credited for constantly viewing the methodology, set-up and experimental design for possible explanations for inconsistencies. They made some effort to understand lack of reproducibility among laboratories. Interlaboratory validation data were presented in a logic manner, making the assessment easy. In the final interpretation an improved logic interpretation progression was proposed and the summary data presented according to this proposal. This was helpful since the opportunity to tests the proposed scheme was used effectively. In some of the inter-laboratory data-sets, one got the feeling in spite of inconsistencies or low reproducibility/ repeatability the final conclusion was clear-cut. So it was difficult to comprehend what the threshold (within the weight of evidence perspective) was for making the particular decision. But, overall seen, the data interpretation was sufficiently clear.

Richard Wassersug: I have concerns about the comprehensiveness and consistency about the interpretation of the data from the various Phase 1, 2, and 3 trials. Almost all of my concerns center on the biology of the anuran larvae and the precision in which the assays were executed in the various labs. These concerns, as they arise in the "AMA Integrated Summary Report" of October 16, 2007, and are listed in order below.

Section 2.1—The first paragraph on the purpose of the assay states that "amphibian metamorphosis provides a well-studied thyroid-dependent process." In truth this has been studied in less than a dozen genera, out of the nearly 400 genera currently recognized in the Amphibia. I think it is important that the EPA documentation for the AMA include a introductory statement on the phylogenetic distance of the genus *Xenopus* from most other anurans, although that is touched upon in one of the background documents [notably ENV/JM/MONO(2004)17].

There is a tendency for molecular biologists, endocrinologists, and toxicologists to believe that what is true of *Xenopus* is true of *Rana*, and that what is true of both of those genera is true for all anurans. This presumption, for instance, is implicit in the introduction to Yun-Bo Shi's 2000 book Amphibian Metamorphosis and in the recent review by Fort et al. in Critical Reviews in Toxicology. In contrast, there are data indicating a variety of ways that *Xenopus* and *Rana* tadpoles differ.

One that may be particularly important to the AMA is how injured tails of these tadpoles respond to a retinoic acid challenge. Most anurans will start to differentiate hind legs and pelvic girdle at this caudal site of injury. *Xenopus*, however, does not show this response.

I would like comment on the appropriateness of *Xenopus laevis* as a model species for anuran metamorphosis assay. The various documents, particularly ENV/JM/MONO(2004)17, review the pros and cons of using *Xenopus*, particularly *X. laevis*, as the model species in the AMA. But they miss a few important points.

X. laevis has managed to establish itself as a feral exotic species on several continents outside of Africa, and can now be found in various disturbed and natural environments (e.g., in California, England, Chile) far beyond its normal range. As such the species appears to be exceptionally robust and tolerant of chemical stressors
(see http://www.columbia.edu/itc/cerc/danoff-burg/invasion_bio/inv_spp_summ/xenopus_laevis.htm). Thus, as acknowledged in the EPA and VOECD documents, a negative response from the AMA with *X. laevis* does not mean that a particular agent does not have a detrimental effect on other Anura. Several studies with the agents used in the Phase 1, 2, or 3 trials, which have yielded a positive response in *X. laevis*, have failed to do so at comparable doses in other species, or vice versa (e.g., see Ortiz-Santaliestra, 2007).

The value of the AMA is obvious when one is exploring compounds of a retinoid nature (cf. Gardiner et al., 2003; Fort et al., 2007). But what seems to be too often either hidden or forgotten is that how the *Xenopus* response to retinoids is quite different than that of many other anurans. This is partially recognized in Degitz et al. (2000), but not in Degitz et al. (2003).

Tadpoles of many species will have a homeotic transformation of their tail tips into limbs, if the tail is injured and then challenged with retinoids. As Maden (1993) points out, this does not happen with *Xenopus* and I have personally confirmed that. I witnessed, however, that *Xenopus* growth was greatly retarded with retinoic acid. With high concentrations of retinoic acid, the tadpoles appear to starve to death (see also Degitz et al., 2003).

Xenopus and *Rana* have fundamentally different anatomy, functional morphology, and growth patterns for their tails and this may, in part, account for the different responses they have to the injury and to a retinoic acid challenge (Nishikawa and Wassersug, 1988). *Xenopus* tails continually add myotomes throughout the larval period, whereas *Rana* tails have deterministic growth (Wassersug 1997). As an aside, in a few studies where *Rana* have failed to show the homeotic transformation, I suspect that the stage of the tadpole and the dosage of the retinoic acid were not appropriate to elicit the response (in agreement with Degitz et al., 2000). This does not moot the utility of using the AMA for studying environmental retinoids. But it serves as a warning on how much one can safely infer from a negative response from *X. laevis* in the AMA.

This does not mean that *Xenopus* is not the best species for the AMA, but it does suggest that more emphasis should be given to encouraging researchers to be ready to explore further when a suspected endocrine disruptor yields a negative assay with *Xenopus*.

Although the concern just raised does not necessarily warrant changing the assay, but it does warrant changing the text. Thus for example, in the second to last paragraph in section 2.1, we are told that "the AMA focuses on anuran metamorphosis because it has been well-characterized." It would be more judicious to be a little more conservative and state that "the AMA focuses on *Xenopus* metamorphosis as a well-characterized example of metamorphosis in the Anura."

Section 3.3—This is the first place where we are introduced to hind limb length as a core endpoint of the AMA. Although this will seem like a trivial point, it may be worth specifying whether the left or right limb should be measured. I doubt that a lateralized difference in the length of the limbs occurs in *Xenopus*, but that is not impossible, considering that handedness in humans can affect aspects of limb size and development, and anurans do show handedness in hindlimb use (Robins et al., 1998).

Another point to consider is whether the AMA should request that both limbs be measured since increased fluctuating asymmetry is a well-established indicator of stress and disturbance during development in many vertebrates (http://www.animalbehavioronline.com/fa.html). Furthermore fluctuating asymmetry has been documented in the anuran appendicular skeleton (e.g., Vershinin et al., 2007; Söderman et al., 2007).

Section 3.4—The paragraph at the beginning of this section does not specify the size of the tanks in the various trials nor whether the flow-through system has the tanks in parallel or series. Granted, in the full description of the AMA both tank size and the flow-through path are specified, but it should be in the Summary as well.

Section 3.6—A key sentence in this section says that compounds which "are thyroid inactive will not likely undergo further testing to characterize thyroid activity." My concern here is that iodinase activity can occur even in frog embryos and can affect nervous system development (Dubois et al., 2006). Hence, anuran embryos can have a component of thyroid activity even before they actually have a thyroid gland. This

suggests that thyroid function can be disrupted in an anuran even without a thyroid gland! This caveat suggests that the FETAX assay should be considered and possibly run concurrently in certain cases where the AMA is also being used.

[As a small note, for some strange reason in this section, and in many tables the abbreviation "HHL" is used for hindlimb length. Elsewhere the more logical abbreviation "HLL" is used.]

Section 4.1—The rationale for using live brine shrimp as a food for *Xenopus* tadpoles in the US labs is not provided. *Xenopus* tadpoles are obligatory microphagous suspension feeders (Seale et al., 1982). It would be interesting to know whether the lab that fed its tadpoles live brine shrimp had evidence that the brine shrimp were effectively digested. If that diet improved tadpole growth, might it have been due to the addition of salt to the water along with the brine shrimp, rather than the shrimp themselves?

In the same section it is mentioned that "test vessel size and tank dimensions were not reported." Could the labs be contacted and asked for that missing information? To accept such important information as 'missing' is problematic. There is reason to believe that for a social schooling taxon like *X. laevis*, the number of tadpoles per volume can affect the growth rates, even when the food is abundant (see Katz et al., 1981). Thus without information on the density of the individuals (and not just the numbers per tank), it is not possible to fully interpret the different results among the different labs.

There is a problem with the anatomical terminology in the section where the thyroid gland histology is described. Here we are told that "transverse sections of the lower jaw" were made. If one is precise about the language, then none of the labs that did that would have found any thyroid issue. This is simply because the thyroid glands in *Xenopus* tadpoles lie within the brachial baskets, caudal to the "lower jaw," as shown in Fig. 1 in the "Guidance Document on Amphibian Histology Part 2." Obviously the various labs took not just the lower jaw, but the whole buccal floor and part of the pharyngeal (branchial) baskets; i.e., the floor of the mouth and the throat. This may seem like a petty point of language, but since there are few pictures in the literature about where the thyroid glands actually lie in *Xenopus* and other tadpoles, it would be helpful in the EPA documents were precise in terms of their terminology about the anatomical location of the gland.

On page 27, just below table 4-2, we come to the first mention of the use of static versus flow-through systems. Here we run into what I consider the first serious problem with the AMA methodology.

The statement on that page acknowledges that tadpole development under static conditions could be greater than tadpoles raised in a flow-through system, even when the same amount of food is provided. This is not surprising since *Xenopus* lives in still water in the wild and, as documented below, tadpoles are stressed when raised in a current. The response of *X. laevis* larvae to currents was examined more than a quarter of a century ago, but that literature is ignored in all of the AMA documents.

The closest the background literature comes to acknowledging the problem is on page 68 of the ENV/JM/MONO(2007)23 document. There it states that possible problems with the "established flow-through exposure system [in the Japanese lab]…may explain some of the slight differences [in results] in the control animal performance." Those "slight differences," though, were the greatest in the inter-laboratory comparisons.

Section 5.1—Elsewhere the potential problem of currents for adults is recognized. So, for example, we find on page 52 under Section 5.111, paragraph 92, the statement: "Since *Xenopus* live naturally in static environments, care is required when using the flow-through systems so that the flow does not disturb the frogs." Since adults are negatively buoyant, benthic, and of relatively large mass, they can easily resist displacement in a gentle current without exerting energy. The tadpoles are not so lucky. Because of their

neutral or positive buoyancy, pelagic life-style, and small mass, they cannot avoid being displaced by a current without expending energy swimming upstream.

The stress to tadpoles raised in currents has recently been investigated in a stream-associated species *Rana boyii*. Dr. Sarah Kupferberg (Questa Engineering, Richmond CA, skupferberg@pacbell.net) has unpublished data that *R. boyii* tadpoles, which are far better designed for handling currents than *X. laevis*, exhaust in a matter of minutes in a current of just 5 cm/s.

Both in terms of the phase trials that were undertaken and the final AMA protocol itself, I strongly encourage the EPA to include document that raising *X. laevis* tadpoles in a current has an inconsequential effects on their growth and development. If the highest concern of the AMA methodology is to provide a continuous dose of the chemicals being assessed, then a rationale should be provided for why that is of higher priority than trying to raise the tadpoles in a slightly more natural and less stressful (i.e., in a non-flow-through) system. If a flow-through system is absolutely required, then much greater detail needs to be provided about the position of the inflow aperture(s) and whether it (they) induce a standing circulation in the tanks. In the current version of the AMA methodology, there is inadequate information on the permissible flow velocity in the tanks. Do the tadpoles line up with their nose pointing towards the inflow? If so, they are showing a positive rheotropic response and will be swimming harder (and expending more energy) than they would be in a static system. In addition, almost any current will cause major, non-random distribution of suspended food particles (see Walks, 2007). How will that affect growth rates and the variance in growth rates for the tadpoles in a single tank?

I do not wish to see this issue delay putting the AMA online as an approved EPA assay. But I do not feel that the AMA methodology can be considered in final form until there is some hard data showing that the inflow current is not affecting the tadpoles' behavior and growth. Since there is no detail provided on the currents generated by the flow-through system in the various Phase 1, 2, and 3 trials, this reviewer does not know whether the variance in the results from the different labs is not largely due to inadequate control of that particular variable.

On page 29, we learn that different labs anesthetized the animals different numbers of time. The final AMA protocol recognizes this as stressful for the tadpoles. None of the phase trials, though, explore this potential variable.

Lastly, no information is provided on the O_2 concentrations in the tanks. So, again, we don't know what importance differences in water chemistry may have had that could account for the different results between the different labs.

The next major problems all center on how one recognizes overt distress in a *Xenopus* tadpole.

Several of the tables that summarize the results from the various tests have a section titled "Overt Toxicity." There are three variables of 'toxicity' listed in those tables that are not strictly morphological markers. These are 'abnormal behavior,' 'lethargy,' and 'reduced food consumption.' To a non-behaviorist, it would seem that none of the labs witnessed any problems at any time in terms of any of these variables. However, since there is no discussion about what is normal behavior for a *Xenopus* tadpole I doubt that many (any) of the labs attempted to assess those variables…or were fully aware of what to look for in terms of behavioral disturbance.

Let's consider first 'abnormal behaviors.' *Xenopus* tadpoles are obligatory air breathers (e.g., Wassersug and Murphy, 1987; Pronych and Wassersug, 1994, plus older literature cited therein). They may come up

to the surface in normoxic water only two or three times an hour, but, if they are stressed, they reduce their aerial respiratory rates. I have anecdotally noted (Wassersug, 1996) that simply tapping on the side of *Xenopus* aquaria can reduce the tadpoles' aerial respiratory rates for up to an hour. Suppression of activity and reduced aerial respiration are well documented in the literature for stressed tadpoles, but never mentioned in the AMA documents.

Since labs that did all of the phase trials do not discuss the procedures they undertook to reduce the stress on the tadpoles, my guess is that all of the tadpoles were somewhat stressed. The problems then, are, "How much?" and "Was it the same amount of stress in all labs?"

Let's take a look at specific *X. laevis* behaviors. *Xenopus* tadpoles normally swim at an approximate 45º angle in the water column. However, if they are in a current they reduce their lung use and lung volume. They then have a shift in their center of buoyancy and swim more horizontally. None of the labs reported on the angle or orientation of the tadpoles. So we cannot tell whether their swimming was "normal," as it would be in standing water, or "abnormal" as it would be if they were swimming against a current and had reduced lung volumes.

When *Xenopus* tadpoles swim faster, they incorporate more of their tail in generating a propulsive wave. However, the frequency of the tail beat changes very little at low to moderate speeds (Hoff and Wassersug, 1986; Wassersug, 1989). What then was the wave pattern in the tails of these tadpoles? No data are provided.

To simply say that the tadpoles were swimming normally and "not lethargic," because the tails were constantly waving, presumes that the tail beat is under neuronal control. *Xenopus* tadpole tail tips, however, can continue to beat in tissue culture media for hours to days. So simply to witness that the animals swimming does not mean that they had normal behavior, i.e., that they were not "lethargic."

What other behaviors might have been examined and scored to document abnormal behavior, stress, or distress? The buccal pumping rate would be an obvious one. But there is no evidence that any of the labs measured this. This, in turn, directs our attention to the other measure of overt toxicity; i.e., "reduced food consumption." How did the labs measure the rate of "food consumption" to know if it was normal or reduced? *Xenopus* tadpoles reduce their buccal pumping rate when in a suspension with a high concentration of food particles (Seale and Wassersug, 1979; Seale et al., 1982). This is understood to be an adaptive response that helps the tadpoles avoid clogging their suspension feeding mechanism (Wassersug and Murphy, 1987). The Phase 1 and 2 laboratories, then, might have measured buccal pumping rates as an indirect proxy of feeding activity. However there is no evidence any lab collected such behavioral data.

A more direct measure of food consumption is a change in particle concentration in the water column around the tadpoles. This can be measured directly with a cell counting system, such as an automatic particle counter, or the old fashioned way using a grided slide under the microscope. But, once again, there is no evidence that any of the labs actually measured changes in particulate matter in the water, so it is not clear how they could have concluded that 'reduced food consumption' did not take place (other than indirectly from the final size of the tadpoles).

Whereas it is only a matter of history to criticize what was or wasn't done in the various labs, what really matters now is what is going to be considered normal tadpole behavior for the AMA. If the AMA is going to include measures of 'overt toxicity' that include behavior, then there must be rigorous and clear guidelines about what behaviors should be observed, how they should be quantified, and what is considered normal. In many ways this is the biggest weakness in the AMA documentation.

Before leaving this section, there is a sentence on page 34 that is unclear. That is where we are told that "hind limb length measurements were less straightforward due to a heterogenous effect in the Japanese laboratory." What is a "heterogenous effect?"

The last paragraph in section 4.3 concludes—despite all of the undiscussed and uninterpreted variation in results between the labs—that "the model system is relatively robust and not subject to variation as a function of the test protocols employed." I frankly do not see how such a strong statement can be made when there is variation in the test results between the labs in either gross morphology or thyroid histology that remains unexplained.

As we proceed through the document and the reviews of the various trials in the various labs, this same problem re-emerges. Thus we see on page 41 the statement that "the inter-study variability for wet weight of controls was somewhat greater." This raises a suspicion for me that the animals were subjected to different levels of stress in the different labs, but not enough information is provided to determine what those stressors were. As we work our way through the various chemical agents, we get hints of more variance possibly in behavior that is unexplained. On page 46, we are informed of sedative effects from phenobarbitol. But were those effects similar in all the labs?

The statement on page 49 of "a finding contrary to expected" would seem to have warranted some effort to figure out the source(s) of the variance. Yet the source or sources are not explored in these documents.

Section 5.2—One more hint that things were not normal (or at least consistent across labs) even in the controls, is the size range of the *Xenopus*. The average maximum size for *X. laevis* tadpoles in the wild is 80mm (Wager, 1965). The maximum size for tadpoles according to ENV/JM/MONO(2004)17 is 60 mm (page 52), but some of the lab results suggest that control animals are metamorphosing well below that size. It is quite likely then that the laboratory stock that have been used in the various laboratories around the world have been subjected to some substantive artificial selection, as well as the fact that the tadpoles may have been raised under non-optimal conditions.

I maintained a *Xenopus* colony for some 30 years. Over the years I found a tendency, when trying to maintain stock, to keep the first animals that metamorphose after a breeding and discard the extra tadpoles. In a few generations, this can lead to a bias for small individuals that metamorphose at a smaller size. I see nothing in the AMA that discusses how to maintain uniformity, if not 'wild type' in the breeding stock used in the assay. That issue needs to be addressed in the AMA methodology. If it is not addressed, then it belies the key statement in the Introduction to this section that "it is also imperative to refine husbandry methods and other test factors to ensure optimal and consistent performance of controls."

Only two paragraphs later, we are informed that "less than optimal control performance occurred in two experiments during the study." Without any effort to trace down the cause of that sub optimal performance, there is no guarantee that the AMA methodology can be consistently executed.

Section 5.4—I consider the inter-lab variation in tadpole size presented on pages 55 and 56 high. What guarantee do we have that the AMA in the future will perform any more consistenly? My concern repeats itself as we go from one test chemical to another. Thus, in section 5.4, we are told that there were "no signs of overt toxicity" but, as noted above, its not clear that the labs looked for behavioral indicators of stress or toxicity. Considering the fact that T4 <u>is</u> a thyroid hormone, one would hope that the assay could run without the level of variability reported for T4 on pages 61 and 62.

When we learn that laboratory 5 had mortality "due to handling errors" warning lights go off in my head. What were the errors? Were all the animals abused, but only a few of them dying? Were those "handling

errors" isolated and specifically involving only the individuals that actually died? Or were all the tadpoles exposed to those "handling errors" and some of them were hardy enough to survive?

Given the variation reported between the three labs, the last sentence on page 67 is bothersome. We are told that "the strong developmental response was deemed to be sufficient to conclude that the assays successfully detected T4." This seems to me a trivial statement, since we have known for decades that T4 affects the development of *Xenopus* tadpoles. What is so problematic is the begining of that sentence where we learn that "thyroid histopathology as inconsistent between the three labs." Thus, for certain agents in certain labs, histopathology is a powerful aspect of the AMA's ability to discern endocrine disruption. In other labs, histopathology yields inconsistent results. Without chasing down the source of such inconsistencies, we cannot have full faith that the AMA protocol can produce consistent results between different labs.

The problem keeps returning. So, on the bottom of page 70, we learn of an additional difference in results from the various labs for which "the reason has not been determined." By now one has the impression that many months, in many labs, were spent to show the obvious; i.e., that compounds like T4, which are thyroid promoters, accelerate development whereas compounds that have long been known to inhibit metamorphosis, do so in more than one lab. Yes, the AMA works! But not ideally, and not consistently. So I'm left wondering why more effort was not put into trying to identify and resolve the variation reported in the results from the different labs.

The various sections all seem to end with some statement that the assay worked. Thus we are told that the strange development observed in the tadpoles in the iopanoic acid (IOP) studies (with "asynchronous development") simply because it gave a response "can be considered a 'positive' result." Yes, positive. But otherwise uninterpretable.

The last paragraph on page 76 states that the iodine content in the culture water "must be considered." It isn't clear that this was addressed in the earlier phase trials and may be more important than is appreciated in the current AMA methodology.

Section 5 ends with a statement that metamorphosis in *Xenopus laevis* could be used as a "testing tool for thyroid system disruption." While this important concluding statement is in italics, this was clearly known twenty years ago.

Section 6.2—Here is an aside on biology, and not on the assay *per se*. I found it intriguing that estrogen increased the size of the *Xenopus* tadpoles. Adult female *Xenopus* are larger than males. Over the years, I have been occasionally asked if there is some way to tell male from female tadpoles. It would be fun now to go back to the lab and find out whether, all else being equal, female *Xenopus* tadpoles are larger than males at or before metamorphosis. Hayes et al.'s (1993) failure to find any estrogenic effect on *Bufo* larval growth and metamorphosis doesn't moot the question. It is my impression that species, whose size at metamorphosis is closest to their size at first reproduction, are more likely to show differentiation of their gonads at metamorphosis than species that metamorphose at a size well below their reproductive size. Sex difference in size at metamorphosis may thus be most likely to be found in the former rather than the latter group.

Section 7—This section acknowledges that the scientific literature was reviewed up to 2003. It is not clear why the literature wasn't updated for the last three or four years. The literature, though, is updated in Fort et al. (2007).

In section 7.2, we are introduced to *Silurana* (*Xenopus*) *tropicalis* as alternative model species. We are told that it could be used in place of *Xenopus laevis* "with minimal modifications," but those modifications are never specified.

Section 8.1—This section proclaims "The reproducibility of the AMA, for screening purposes, has been well-demonstrated using several representative thyroid-active chemicals across geographically diverse laboratories." However, if the variation between the labs cannot be explained, then one cannot feel as confident about this proclamation as the author of the review.

Section 8.3—Here the strengths and limitations of the assay are listed. I agree with the combination of morphological and histological endpoints, but they are only considered acceptable within the context of the animals having normal behavior. Without defining 'normal behavior,' and without any clear guidance on how to quantify that, it is not clear how sensitive, reproducible, and reliable the AMA will be.

Comment on the Biological and Toxicological Relevance of the Assay as Related to its Stated Purpose

David Crews: This assay is designed as a standard toxicological screen. As such, it accomplishes its goal. However, a number of studies have now shown unequivocally that traditional toxicological studies are ill-suited for detecting chemicals that have endocrine disrupting capacity. There are multiple reasons for this and are listed below.

a. The thyroid system is part of an integrated endocrine system that is essential not only for normal functioning at particular life stages, but also for advancing the developing organism through a series of carefully regulated stages that result in a functional (= reproductive adult). Hence, it cannot be considered in isolation of other endocrine systems. This is particularly the case when considering the developing reproductive system.

The present document describes the effects of compounds on the thyroid axis and its consequences on limb growth. This ignores the fact that factors influencing the thyroid axis may also affect the reproductive axis. For example, a recent study comparing populations in frogs in a contaminated (by agricultural runoff) and a pristine lake in Italy document that the pattern of circulating concentrations of steroid hormones and T3 and T4 are disrupted and the testes of adults affected (Mosconi et al., 2005). In laboratory experiments administration of goiterogens such as thiourea and 6-*n*-propyl-2-thiouracil (**PTU**) can alter normal patterns of sex determination in Xenopus and other frogs as well as fishes and mammals (Fort et al., 2007; Franca et al., 1995; Hayes, 1997a, b; Matta et al., 2002; Schultz et al., 2005). Significantly, after treatment is stopped, spermatogenesis is restored to normal levels (Cooke, 1996; Kirby et al., 1992; Schultz et al., 2005). Thus, such compounds lead to increased interstitial cell growth and activity, to the extent that in the male spermatogenesis is inhibited (presumably due to an overproduction of androgen). Such studies indicate that thyroid hormone is important in normal gonadal development and, further, that interference at this level will produce sterile individuals.

b. EDCs are ubiquitous in natural environments. Standard toxicological screening methods have a focus of determining whether a given compound is toxic, leading to death (defined here by the LC 50 and/or to body and organ malformations). Two typical life stages in which compounds are tested are the adult or developing (embryonic or early life) organism. In both instances the emphasis is on the individual organism within a single generation. In addition, any number of compounds when administered to developing organisms may have no demonstrable effect on mortality or growth. However, these compounds, and particularly EDCs, can affect sexual development—even at extremely low doses. Such sterile individuals occupy space and use resources but cannot contribute to the growth of the population, as their genes will not transmit to subsequent generations, hence leading to evolutionary death.

J. David Furlow: The biological and toxicological relevance is clear: metamorphosis is a strictly thyroid hormone driven event, therefore it is reasonable to assume that alterations in the progression of spontaneous metamorphosis by toxicants are the result of disruption of thyroid hormone synthesis and/or action.

Catherine Propper:
a. In the validation processes it would have been useful to determine whether an exposure protocol from hatching through metamorphosis provided a different outcome than either of the shorter protocols (stage 54/14 day or stage 51/21 day). One of the issues this assay did not answer in the context of the presented validation phases is whether early exposure (prior to stage 51) impacts later thyroid-related outcomes. Animal (including human) populations may be exposed to these compounds throughout their lives, not just from a specific stage on. This early exposure may have impacts on the thyroid system that will not be seen here. For example, we have preliminary data in our lab where we see impacts on timing to metamorphosis from exposure to complex mixes that we do not see with a 14 or 21 day assay (Propper, unpublished data).

b. In this assay validation approach, through 3 phases of the validation process only three known environmental contaminants were evaluated. These are perchlorate, which has well-documented means of thyroid hormone disruption, estradiol which in fact does have some affects on the thyroid hormone system (see comments above), and benzophenone-2. The other compounds tested were chosen largely based on their known pharmacological thyroid hormone disrupting activities.
A final validation step evaluating some environmental contaminants (eg. Pesticides or pharmaceuticals) known to have thyroid disrupting capacity at environmentally relevant concentrations is necessary for final knowledge of the utility of the AMA Test Methodology. For example, endosulfan has demonstrated impacts on thyroid gland histopathology, and impacts thyroid hormone levels in a number of species (including human pesticide formulators). A final validation step demonstrating the usefulness of this assay using a common environmental micropollutant would strengthen the justification of the protocol. Another advantage of including such a validation is that the probability of getting a monotonic dose response is much less, and therefore tests the validity of the assay when a U-shaped (or other non-linear response) is generated.

Hannes van Wyk: In all the literature presented, including earlier DRP and recent published literature, toxicological relevance of assays focusing on environmental (external) thyroid modulation with potential adverse consequences for wildlife and human health, will always be relevant. An extensive literature now exists that suggest a range of environmental toxicants that may in some way interact with the thyroid system. The Introduction and Background sections of all the documents recognize this phenomenon.

As acknowledged by most authors, the general control and endpoint expression associated with the thyroid system is rather complex. More the reason to understand the range of potential mode of actions to ensure toxicological relevance. I am convinced that this point is clearly made in the "Rationale for the assay". The rationale for employing non-mammalian organisms as models for assessing thyroid disruption seems to be convincing and acceptable, especially when considering the recognized evolutionary conservatism among vertebrate groups. The AMA uses the advantages that amphibians offer to study endocrine disruption of the thyroid system through phenotypic thyroid hormone (TH) dependent changes during the developmental phase (metamorphosis). The role of TH during early amphibian development (with free-living embryos) and early mammalian development underlines the relevance of using the AMA, a simpler more straightforward system to work with than working with early mammalian life stages.

The authors clearly and comprehensively reasoned the relevance and advantages of using an anuran metamorphosis model in studying external influences on the thyroid axis. In Section 2.2, they summarize the dynamics of hormonal changes during the developmental programme. Similar changes in expression

of TH-receptors were presented elsewhere. It is clear that during the development, refinement and validation phases of the AMA considerable thought has been given to the relevance of the exposure window. It is also clear that several possibilities exist to use short term, molecular based TH receptor expression along with the longer-type assay using morphological based endpoints. Although, it seems that earlier suggestion for the inclusion of the former did not materialize as integral part of the AMA.

The importance of controlling for several environmental conditions that may secondarily affect the rate of metamorphosis was also shown. This must be valuable to the user of the AMA, specifically to understand the sensitivity of amphibian development to a range of environmental factors and therefore the importance of controlling for these to ensure the correct interpretation of exposure data.

The authors adequately describe the possible points of modulation and uses Figure 2.2 to show the non-neuro-endocrine (or peripheral) points of concern. It is not clear why they selected to omit the potential points of effect on the neuro-endocrine side?

Reading the DRP and the ISR together, I am convinced that the extent of literature review to set the scene and build the rationale for the AMA is extensive and represents a good review of the literature to highlight the hormonal control of amphibian metamorphosis. It has been shown that the AMA represent an opportunity to study several TH dependent endpoints and mode-of-actions rather than just screening for the ability of chemical to bind to TH receptors (like in several HTPS assays). Apart from the classical genomic interactions, non-genomic interactions as well as pathway enzymes involved in synthesis and metabolism activities may also be included. In summary, therefore I am convinced that the biological and toxicological relevance of the AMA has been shown. Although it runs the risk of being too "reductionistic" when it comes to EDC action, it represents a broader multi-endpoint perspective, and therefore, certainly conforms to the goals of a screening assay for suspected/potential Tier 1 EDC interaction.

Richard Wassersug: The AMA with *Xenopus* is toxicologically relevant in that this is the most common amphibian used in toxicological research around the world. Its biological relevance, however, is slightly less relevant in some situations. *Xenopus* is not native to any continent outside Africa, and its morphology, ecology and behavior both as a larva and adult, are quite unlike those of other amphibian genera in North America, Asia, Europe, or Australia. The authors of the EPA documents suggest that the agency is aware of situations where data collected with *Xenopus* via the AMA, may not be relevant for other species and mentions the potential need to verify the results from the AMA with other anuran taxa.

Provide Comments on the Clarity and Conciseness of the Protocol in Describing the Methodology of the Assay such that the Laboratory can a) Comprehend the Objective, b) Conduct the Assay, c) Observe and Measure Prescribed Endpoints, d) Compile and Prepare Data for Statistical Analyses, and e) Report Results

J. David Furlow: Comments on assay method narrative:

a. A major source of uncertainty in my mind lies in the use of flow through versus static renewal systems (p. 2). The static renewal system (for tadpoles) is likely the most popular system in most laboratories working with *Xenopus laevis* tadpoles due to convenience and cost (and it is understood that this system may be the only option in some toxicology studies for chemicals with certain properties). It should be noted that silt-filled, murky ponds are apparently the natural habitat of *Xenopus laevis* rather than fast flowing streams. Furthermore, the flow through system would not permit accumulation of degradates of the test chemical that may occur due to light, hydrolysis, and the animal's own metabolic capacity. Nevertheless, the flow system is

understandably preferable due to the higher degree of reproducibility and better control of water quality. According to the ISR, (p. 20) a particular exposure system is not *required* but that the flow through system is *preferred*. It would be important to take perchlorate, for example, and test the flow through versus static renewal systems in the same laboratory using the same spawn. In summary, without such a direct comparison, either a static renewal system should always be used so all chemicals can be tested, or the flow through system should always be used and just exclude chemicals that are not suitable for the assay conditions.

b. The statement about "suitable plastics" for system components that do not compromise the study is not clear: certain plastics can likely be ruled out right away such as those that leach BPA and other known endocrine disrupting chemicals (p. 2).

c. In addition, in the flow through system as described on p. 2, it is important to specify that each of the four replicated doses receive an independent water supply (rather than from the same source split into four in order to serve as four independent samples). This point was not clear in the assay description. Perhaps a diagram can be included to clarify the system design.

d. The protocol should recommend whether to dejelly the eggs of spawns used for the assay rather than leaving that up to the individual investigator (p. 5). Dejellying allows much easier sorting of poorly developing embryos that may compromise the rest of the batch. Thus a recommendation one way or the other should be made.

e. The assumption is that the chow from Sera Micron is consistent from lot to lot, but how is this assessed? Are there any guidelines on expiration date or storage conditions? (p. 6).

e. For vehicle controls, a range of concentrations of the most common (ethanol, DMSO) can be tested in the system for effects on metamorphosis (or lack thereof) to make recommendations to testing laboratories. (p. 7).

f. The choice of dosing regimen is unclear (p. 8). While the determination of the MTC is basically clear (although the description of other means to estimate the MTC is rather convoluted), I see a problem with allowing only three doses to be tested with a dose separation of 0.33 to 0.5. For example, the example given does not even satisfy the requirements of the assay as stated: 0.11 of the highest nominal concentration of 1.0 is only 1/9 of the maximum dose. The risk here is that that the assay may not be able to discriminate between general toxicity and a more specific effect on the thyroid hormone driven metamorphosis that may be revealed at lower doses.

Richard Wassersug: My greatest concerns about the AMA center on the document "Draft Method for the AMA." Various laboratories should be able to follow the methodology of this essential document and achieve identical results. There is simply not enough detail in this methodology to be confident that the assays can be executed with adequate amounts of reproducibility.

The following is a list of my major concerns.

Breeding stock—No guidance is provided on whether one should be concerned about inbreeding in laboratory stock. As noted in Item 2, the labs in general seem to be reporting tadpoles in control tanks metamorphosing below the maximum size in nature. As I've noted above, it is easy to artificially select for larvae that metamorphose at a small size. But how would that affect the results of the AMA? One guess is that it would reduce the sensitivity of the assay. If a presumed endocrine disruptor reduces the size of tadpoles, and the tadpoles used in that assay have already been selected to be dwarfs, then it's going to be more difficult for the AMA to pick up a significant reduction in size.

I do not recommend that the EPA delay putting the AMA into operation. But ways to either deal with or avoid using inbred lines need to be addressed. Whatever their guidelines are, they have to be tight enough that they yield standardized breeding stocks across various labs.

Exposure system—Another major concern I have is with the mechanics of the flow-through dilutor system. I understand that the tanks will be in parallel, not in series, which, of course is essential. But much more information is needed to make sure that all the labs produce comparable circulation in their tanks by: 1) having identical placements of the inflow and outflow apertures, 2) apertures of identical size, 3) yielding identical flow rates and circulation in the tanks.

In Item 2 above, I emphasize that *Xenopus* tadpoles live in non-flowing water and that putting them in a current is stressful. The background literature in support of this claim goes back at least a decade, some of it to the early 1980s. Nowhere in these documents do I see those concerns mentioned or discussed. Minimally the AMA should include ways to minimize the current velocity in the tanks, such that there will not be a major, standing circulation.

Please consider the following: *Xenopus* tadpoles in a current reduce their aerial respiration rate. They do this by lowering the volume of air in their lungs. This makes them more negatively buoyant so they can stay closer to the bottom, where the flow rate is lowest. This, however, lowers their stamina and can increase their lactic acid concentration (see Wassersug and Feder, 1983; Feder and Wassersug, 1984). If the lactic acid is elevated, then the animals are stressed. Stress increases corticotropin-releasing factor (CRF) which has been shown to activate both adrenal (interrenal) and thyroid hormone secretions (see Denver, 1996, plus other papers cited therem as well as reviewed in Wells, 2007, p. 608 and Fort et al., 2007).

What is remarkable is that the EPA documents fully acknowledge the problem of stress from an endocrinological perspective, yet completely ignore it from an ecological and behavioral perspective. Thus, in the ENV/JM/MONO(2004)17 document (which is in general an excellent document) we are told explicitly on page 27 that CRF, not TRH (=thyrotropin releasing hormone not "thyroid receptor element" as claimed in Table 1-1, page 20 of the same document) "is the primary hypothalamic releasing hormone responsible ultimately for the induction of metamorphosis." On the next page we learn that many tissues in tadpoles are responsive to the impact of corticoids on thyroid hormone action. The section ends (paragraph 29) with the statement that "Overall, physiological synthesis and secretion of corticoids play an important role in anuran metamorphosis." In layman's terms, these quotes recognize that the endocrinological pathways that respond to environmental stress interact with the endocrinological pathway that control metamorphosis. Yet the AMA documentation says nothing about how to limit, or even recognize and regulate non-chemical environmental stressors on tadpoles.

Since the EPA is committed to a flow-through system, in order to stabilize the delivery of the test compounds, far more effort needs to be spent on how to do this in a way that minimizes—or at least standardizes—the stress that currents, for example, place on *Xenopus* tadpoles.

Removing the jelly from the eggs—An optional step in the production of tadpoles for the AMA is to use L-cysteine to remove the jelly. It is not clear why this should be done, optionally or otherwise. From a historical perspective, one can understand why many labs do this. It is, for example, part of the FETAX, which is an assay for developmental disruptors of embryogenesis. Since the concern in that assay is to get the test agent to the embryo in a consistent fashion, it makes sense to remove the jelly, which may or may not be uniform on different eggs and may inhibit transfer of the test chemicals to the embryos themselves. Removing the jelly is also a step in all trangenic work with *Xenopus* eggs. However, in light of the

concerns that iodine in the water may be an important variable that needs to be controlled, I feel that the L-cysteine step should not be optional.

A case can be made for removing the jelly to make sure that iodine and other growth promoting elements in the water (most notably O_2) are not blocked from getting to the embryo. Notably, this has relevance to the 'thyroid axis' even in the early embryo. Dubois et al. (2006) point out that thyroid hormone is assumed to be absent in embryos before they develop a differentiated thyroid gland. However, they show that elements of thyroid hormone signaling pathways are present during early development of *Xenopus*. They find, for example, functional deiodinase activity and even T4 at significant levels during early embryogenesis, this pre-thyroid gland hormonal activity is substantive in neurogenic areas.

An implication of the Dubois et al. study is that thyroid hormonal function can affect tadpole development long before the tadpoles reach NF stage 51. Without more knowledge about how the jelly affects this embryo biochemistry, a case can be made for removing it from all eggs to strive for better consistency. [Minimally, those who run the AMA need to have control of iodine concentration in the water right from the time that they start breeding the adults, and not just during the execution of the AMA.]

There is, however, an alternative way of looking at this. If we are concerned about whether a certain agent is an endocrine disruptor in the natural environment, we should remember that frogs' eggs all have gelatinous coats in the wild, and this material may have a protective function for the embryos. If the results from the AMA are to be most meaningful for other species in the wild, a case could be made for leaving the jelly on, to help make the *Xenopus* eggs more comparable to those of other species in the wild.

Either way—with or without jelly—the EPA should arrive at a consistent and non-optional policy about how the eggs for the AMA should be raised.

Larval care and selection—The AMA similarly must come up with clearer guidelines on how to standardize, if not minimize, the daily disturbance to the tadpoles. In the Methods document there is only a single sentence on cleaning the tanks. There we are told that the tanks "shall be siphoned clean daily." There are no guidelines on how to do this in a standardized fashion that minimizes the stress on the tadpoles.

As mentioned above, tapping on the side of an aquarium can cause *Xenopus* tadpoles to reduce their aerial respiration rate, even when their swimming and other behaviors appear perfectly normal. Siphoning the bottom of a tadpole tank must surely be a comparable or more extreme stressor.

It is well known for tadpoles of other species that they retreat to the shallows and stay near the bottom when they sense a threat. Clearly, intensively siphoning the tank would be a stressful mechanical disturbance for any tadpole. Rot-Nikcevic et al. (2005) found that mechanical disturbance can indeed reduce the growth rate of *Xenopus* tadpoles. Although their data were not statistically significant at the $P < 0.05$ level, their mechanically disturbed *Xenopus* tadpoles were on average 10% smaller than undisturbed tadpoles.

Older data in Wassersug and Murphy (1987) show that aerial respiration facilitates growth in *Xenopus* larvae. Denying *Xenopus* access to air by stressing them so they avoid the air-water interface is likely to retard metamorphosis (Pronych and Wassersug, 1994). Feder and Wassersug (1984) show that 16.6% of the total O_2 consumption for *Xenopus* larvae in normoxic water comes from aerial respiration. This can increase to 100% in hypoxic water. All of these data suggest that mechanical disturbance is likely to negatively impact on *Xenopus* larvae in the AMA. This mechanical disturbance can be from cleaning activity, noise from pumps, human activity around the tanks, bubble stones or other aerating machinery,

etc. In order for the AMA to yield consistent results between labs, the protocol must include rigorous standards for controlling, if not eliminating, these sources of stress to the tadpoles.

Establishing the highest test concentration—There is a subtle contradiction in the example given under the subheading of "test concentration range." There we are told that the minimal range "shall be at least one order of magnitude" but that is immediately followed by an example where the range runs from 0.11 to 1.0, which is slightly less than one full order or magnitude.

Daily observations of test animals—We are told this is necessary, but there are no directions about what one should be observing. Yet again, it seems imperative that the AMA define more rigorously what constitutes normal behavior for *Xenopus* tadpoles.

Hindlimb length—Should the same side of the tadpole be measured in all the labs? Should labs measure both sides so they can collect data on fluctuating asymmetry?

Body length and wet weight—More direction is necessary to standardize how one should remove adherent water from the body of tadpoles before their weight is determined. The document recognizes that "weight determinations can cause stressful conditions for tadpoles and may cause skin damage." This would mandate standardization in this step. Over the years I've watched students very gently pick up tadpoles with a dipnet and do virtually nothing to remove surface of water for fear of injuring the larvae. I've also seen tadpoles get shaken down vigorously and patted dry as if they were vegetables being prepared for a salad. The EPA needs to provide greater direction about how the tadpoles should be freed from surface fluid in order to increase the chances of comparable weight measurements between labs.

Additional observations—The text here makes it clear that the EPA expects behavior to be monitored, but it gives no guidance on how to do this. Taking each one of their examples, one can see problems.

They start off by mentioning "uncoordinated swimming." *Xenopus* is a social species. Is "uncoordinated swimming" then measured by the geometry of the school (e.g., orientation of one tadpole to another? distance between tadpoles? etc.). The distance between tadpoles varies depending on their size, density, and illumination (Katz et al., 1981). But chemical agents can also affect the interactive distance; i.e., the 'coordinated' nature of their swimming within a school (Lum et al., 1982). Should this be measured to determine if their swimming is coordinated?

The next variable mentioned is 'hyperventilation." Ventilation for *Xenopus* tadpoles has both an aerial and an aquatic component. Under normoxic conditions, tadpoles come to the surface to take air about twice an hour. If they were to come up three or four times an hour, that would be a 50 and 100% increase in their aerial respiratory rate and could be considered "hyperventilation."

One may suppose that the authors of the AMA protocol were not thinking about aerial respiration at all, but only aquatic ventilation. There is, however, still a problem. The primary determinant of buccal pumping rate (i.e., aquatic ventilation) is not O_2 concentration, but the density of particulate matter in the water (see Feder et al., 1984; Seale et al., 1982). Thus a "hyperventilating tadpole" may be experiencing hunger rather than respiratory distress. Without standardizing exactly when food is delivered to the tanks, how uniformly it is dispersed in the water, and how rigorously ventilation is measured, there will be no way for any lab to determine whether the tadpoles are indeed hyperventilating.

Next on the list is "atypical quiescence." I have no idea what that means or how it is supposed to be measured.

The last variable is "non-feeding," but again there is no indication of how that is supposed to be measured. *Xenopus* tadpoles can regularly feed on suspended particles that are too small to be seen with the naked eye; they are continuous, obligatory, suspension feeders. If they were not trapping particles in mucous, the particles would be going into the mouth through their gill slits and out again. They would then be "non-feeding." But how would any lab determine that?

Possibly the author(s) of the AMA protocol expect those using the AMA to be measuring buccal pumping rate. That is the only variable which can be easily measured that is an indirect behavioral proxy for whether a tadpole is feeding or not. But there are no guidelines provided about how and when to do this.

O_2 concentration—The AMA sets a range for O_2 concentration which should be no less than 40% of air saturation. It does not specify how the water should be aerated in order to maintain that concentration. That needs to be standardized in order to reduce disturbance to the tadpoles.

Water temperature—The water temperature is supposed to be maintained at 22 +/– 1 ºC. This is slightly above preferred room temperature for North America, which is usually 21ºC. With air temperature of precisely 21ºC, evaporative cooling would lower the water temperature to slightly below the 22 +/– 1 ºC range. That would then require some way of heating the water to bring it up to 22+/- 1 ºC. How is that temperature supposed to be maintained?

There are various options for maintaining the tank temperature above room temperature. They range from individual heaters in the tanks to heating the water in the up-stream reservoir for the flow-through system.

I did not see documentation on how different the growth would be for the *Xenopus* tadpoles, if they were raised, say, at 21.1 versus 22.9 ºC, even though both would be in the acceptable range of 22 +/– 1 ºC. It is not clear how the range of +/- 1 ºC was established. One suspects that it was simply convenient and not based on firm data to show that there were no differences in the growth and metamorphosis of *Xenopus* at 21.1 ºC versus 22.9 ºC. In a flow-through system, it can be difficult to maintain thermal constancy within a tank. More guidance should be provided about how to stabilize the temperature in the tanks.

Comprehend the Objective

David Crews: objectives stated in AMA Test Method and Appendices (File names: AMA_Test_Method, Appendix_1; and Appendix_2) were clear and concise. Table 3 should also include daily observation of gross morphological deformities to be consistent with text (File name: AMA_Test_Method, pg. 8)

Catherine Propper: The objective as stated in the test methodology is very short and to the point; however, it would be useful to provide references or weblinks to the other documents that were provided to the peer reviewers so that the labs conducting the tests have access to all the justification for the development of the assay.

Hannes van Wyk: The AMA is structured in such a way that the laboratory should be able to comprehend the objective of the tests to eventually answer the questions related to the purpose of the assay. The selection of *Xenopus laevis* as the test species is explained in the ISR as well as in the DRP. In the DRP comparisons are made between potential test species. From all this it seems that *X. laevis* is still the appropriate species to choose. One aspect of concern is the fact that hCG is used to initiated breeding in captive populations. Very little information on the potential effects of hCG on the response of the thyroid axis to external compounds are available. This is especially concerning when considering the dose of hCG used. Although the AMA will be used to screen chemical compounds and hopefully also mixtures of compounds, therefore, in laboratory studies, the use of local endemic species will have the added

advantage of answering environmental questions. However the fact that *X. laevis* is fully aquatic makes the exposure protocol simple. Table 5.2 seems to be a good summary of comparisons among different candidate species. (The reference to *X. tropicalis* as a South African clawed frog is incorrect, West African?). In summary, enough evidence are available that suggest that *X. laevis* is a robust model and currently the best amphibian species available suited for use in the AMA, with several advantages in handling and breeding of tadpoles for in-laboratory exposures. However it may well be that several other amphibian species could also be used to answer specific questions regarding thyroid endpoints. The knowledge explosion regarding *X. laevis* clearly makes it a valuable aquatic indicator species. Models, like *X. tropicalis* and other local endemic species, may in future be used to answer specific questions, but in the mean time *X laevis* seems to be the best studied non-mammalian model to study aspects of thyroid functionality.

Conduct the Assay

David Crews: Methods and materials in the documents mentioned above were detailed.

Catherine Propper:
a. In general, the test method is missing several details that are necessary. First, there was interlaboratory variation in the validation phases of the test methodology development. To minimize such variation, the assay methodology must be very clear and detailed with acceptable alternatives to the specific methodology clearly delineated (as well as unacceptable alternatives). Such detail is necessary to insure 1) that there is consistency in approach among any EPA contracted laboratories, and 2) that there is consistency in use of the assay by non-EPA researchers who are trying to adopt this assay to their labs' specific hypotheses. Specifics are addressed in the context of the specific heading within the AMA Test Method document.

b. Exposure System: The exposure methodology needs more details in several areas outlined below.

1. The flow-through system is designated as the system of choice, but an option is provided for static renewal. There are problems associated with the justification of the choice and with the description of the methods for using either of the choices.

a. Static renewal: If static renewal is to be used then details of how the water is removed or how the animals are to be transferred to clean tanks needs to be carefully addressed. Then if tanks are to be reused, methods for cleaning and rinsing all the glassware between water changes also need to be provided.
The AMA Test Method protocol states that a complete water change is made if the static renewal system is to be employed. This method implies that the animals must be moved which can cause stress and damage to the animals. A complete water change also removes any bacterial communities that have developed in the tanks that may be necessary for appropriate tadpole development (although if complete water changes were used in the German lab, then it may not represent a problem as the controls performed similarly to the other labs using the flow through system).

I have searched all of the provided documents, including the methods for the Phase I trail and in the "Annex" of the Phase I trail report for the details of the German lab's methodology for static renewal. There is no description of how the static renewal was conducted (and in fact in the "annex," the method is referred to as "semi-static." What does "semi-static" mean?). The provided AMA Test Method provides very few details except that there needs to be a complete water change at least once every 72 hours, and every 24 hours if justified by criteria that are not well defined in the document. A whole water change every 24 hours will be extremely stressful for the animals, and since stress and thyroid interact in this species to impact developmental timing, such frequent water changes must be avoided. If contracted labs are allowed to use the static renewal approach, much more detail needs to be provided in Final AMA Test

Method document, including handling of the animals between water changes, whether the entire volume of the water is changed, and how the animals are to be dosed. Also, whether all the replicates are to be refilled from a common water source with the exposure chemical diluted, or is each replicate dosed independently, needs to be considered.

b. Flow-through system: Again, more details need to be provided. First, does each replicate tank receive an independent water source made up by independent dilutions of the stock solution, or do they come from a common water source (I recommend the former to maintain replicate independence). Second, one type of plastic tubing is recommended in the AMA Draft, but the method states that other unspecified types are acceptable. It is absolutely critical that both acceptable and *unacceptable* plastic tubing be listed. The method needs to specify that the supply tanks must also be glass, and how often the supply tanks are refilled needs to be specified as well. For example, should the tank be refilled daily, every other day (clearly a larger volume will need to be made from stock), or weekly? For pumping the water from the supply tanks to the exposure tanks, more detail would be helpful. Getting exactly 25 mls/min via gravity feed is not easy, and making sure each tank gets exactly the same flow rate would be very difficult indeed. Inexpensive pumps that can be set for such a flow rate should be recommended.

2. Adult Care and Breeding: Consistency in the breeding protocol needs to be strong, and the detailed methodology should be provided here and not just referred to an unreferenced FETAX methodology. Also, it needs to be made clear that using older frogs can lead to delayed development in the tadpoles. The breeding frogs should be purchased for breeding not more than a year before the study. This information is buried deep in *Xenopus* breeding information available online, but I have personal lab experience to attest to the fact that older animals produce slower developing larvae.

3. Larval Care and Selection pages 5-6 AMA Attachment A1. This section needs much more detail.

a. Using tadpoles from one spawn is insufficient. If animals from only one pair of breeding animals is used, any effects (or lack of effect) from exposure found may be strictly due to the sensitivity (or lack of sensitivity) of the one pair's offspring. Three spawns from three separate breeding pairs are really the ideal. Equal numbers of animals from each spawn can be distributed among the tanks. It may increase the variation slightly, but it avoids the risk of pseudoreplication based on a sample size of 1 spawn. In the mammalian literature, peer-review would never accept data supplied from the treatment of 1 litter alone.

b. Is the 2% cysteine placed in the breeding media or in the culture media?

c. *What is the culture media for raising the hatchlings and what is the culture media for rearing during the exposure? This detail is critically important*. In the Phase trials there were some differences among controls suggesting that the media may be important. For consistency, one type of control media should be recommended and made up from preferably deionized or even e-pure water that has the salts (including iodide) added back. Experience from my lab precludes dechlorinated charcoal filtered tap water (there is still something in that water that is toxic to our animals). Other labs may find similar problems. There are several potential options that would lead to consistency in growth media. Labs should either use FETAX (very unpopular among some researchers I have communicated with, but still used by others), 10% Holtfreter's media or some other modified water with salts added back (some *Xenopus* supply outlets even provide their own salts), *but one version should be chosen for the AMA Test Methodology, and it should not be region-specific tap water.*

If the culture water for rearing is the same as for exposure, it needs to be explicitly stated. If it is to change, for example from FETAX to some other media, that needs to be noted, and again, one type of water (not regional tap water) needs to be chosen. Also, once exposure starts, should the exposure tanks receive the water for a specific amount of time before transferring the tadpoles to the tanks? Last, I would

recommend that the tanks must all be aerated during the exposure period and that the DO is measured daily in all tanks and noted.

d. What is the density of the animals in the hatching tanks? What is the volume of the tanks, what is the volume of the culture media in the hatching tanks? All of this methodology should be provided.

e. Are the clutches from each spawn mixed in the hatching tanks (they should be, but if not, they need to be evenly divided within each replicate for all treatments: see comment 3a above)?

f. Under "Larval care and selection," the Table 2 on page 6 should be clearly referenced.

g. The Pre-exposure protocol, page 5 needs more detail. If this pre-exposure period is supposed to provide conditions similar to those of the exposure period, then 1) Static renewal should only be used if it is to be used in the exposure system to, and 2) the flow-through rate should be the same as in the exposure period (25ml/min).

h. Is the water volume reduced once the 5 tadpoles are removed on day 7?

4. Dosing:
a. Analytical Chemical Sampling page 6 AMA Attachment A1: It needs to be made very clear that the quantification of the exposure chemical is to be done for each replicate not just for a representative replicate. In the flow-through system, it also should be made clear that the supply tanks be measured at least once at the beginning and once at the end of the experiment. Details for how much water is to be removed or for determination of such for the chemical quantification needs to be supplied. Further, how the samples are to be stored needs to be provided. It may be that for new compounds such information is limited, but guidelines need to be developed and provided for this methodology.

b. Dose Determination page 7: The issue of dosing is very complicated, and the basis for the decision making outcome is not adequately addressed in the AMA Test Method or in the other documents. The decision to start at a dose that is at the maximum tolerance level (10% mortality) or 100 mg/L, whichever is lowest, has little justification based on the endocrine disruption literature. This level can be at the 100 parts per million range which can be anywhere from 10,000 to 1,000,000 fold greater than is often seen for endocrine disruption. Furthermore, such dosing would potentially lead to compounds being tested at ranges that would far exceed their levels in the environment. Given that: first, many studies have shown thyroid disrupting effects at levels well below these recommended exposure levels; second, the impacts on the endocrine system often do not show a clear linear dose response; and third, this level of testing does not take into account the potential levels of the compounds in water or sediment, how will the results be interpreted in a regulatory environment given that no effect level may not be found with the minimum exposure dose being potentially 11 mg/L?

5. Attachment A2: Embedding tissues. There is one inconsistency: Part 9. States that the head is oriented either ventral to dorsal, ventral side down or "rostral to caudal" and then "caudal side down." To be consistent, need to state that the head is oriented "caudal to rostral" caudal side being the leading edge of the block.

6. Attachment A2: Sectioning tissues: Part 4J, page 9. This section is critical and therefore needs more detail. It would help to state at the beginning how many final sections are to be mounted and stained, and about where in the tissues these sections are to be collected. Having done a lot of histology, it is possible for me to take a best estimate of what is suggested by this methodology, but the step sectioning and examination of the sections prior deciding which to finally mount is not written very clearly.

Hannes van Wyk: ***Breeding of Tadpoles:*** As mentioned before I have a concern about the use of hCG in general but secondly the dose applied seem rather high. Successful breeding and tadpole production can be obtained with much lower concentrations. Although the higher dose ensures large number of tadpoles, the question of secondary effects comes into play. The question of seasonality may be a problem if the laboratory received recently collected frogs from South Africa. Using frogs collected from natural sources for breeding purposes show some seasonality in terms of response to hCG stimulation and egg production. Whether this response is lost with acclimatization and after what period of acclimatization is not known to me.

Following spawning, the SOP states that that the best spawns should be retained. This decision is based on embryo viability. How is this determined? Hatched embryos should be removes as soon after hatching a possible since the water quality goes bad soon after hatching because of all the unhatched eggs. Not convince that the cysteine treatment is necessary. Also not sure about the pipet collecting method. The suction action of the pipet may impact on the embryo. Netting free swimming hatchlings with a flat scoop net seems better. Density control is important during development.

Staging of tadpoles: Although Nieuwkoop & Faber (NF) staging is not too difficult, the criteria used to stage the tadpoles are not clearly stated. I feel more effort should be made to describe the characters to be used (or show visually). Size (WBL) may be variable. N&F state that the optimal size at NF stage 51 is 28-36mm but the Appendix A1 give a range of 24-28mm. NF stage 51 describes the forelimb as oval vs conical in Stage 52, the hindlimb as conicle in shape and the length of the hindlimb as 1.5X its breadth. For a newcomer the staging may be difficult and more detailed or clearer description of the important stages are needed, in particular for the landmark traits. A table summarizing these landmark traits with a pictorial guide will ensure more accurate staging. Stages 51-57 are based on the growth of the hind- and fore-limbs. Stage 58 states that the forelimb is free from the atrium (a landmark). Then criteria switch to aspects of the forelimb (length to hindlimb). Is this how EPA is using the staging? N&F include detailed descriptions of all organ development. The question is which of these can be used confidently by persons doing the staging? Standardization of criteria used, otherwise more variation

Selection of exposure period and length of exposure: The selection of the exposure window did get some consideration during the refinement stage of the assay. Clearly this aspect got considerable thought, discussion and testing in the end. It therefore seems acceptable to use the assay in the suggested window for a period of 21 days.

Exposure procedures (set-up):
Very limited information or reference to the protocols suggested to dissolve chemical in treatment water was given, in particular the liquid-liquid and glass wool saturation systems. If carrier controls are included?

Flow-through vs semi-static. Although in the initial descriptions of the exposure system, the semi-static renewal systems was described in detail limited information is given about the design of the flow-through system. Flow should be low since in its natural environment, *X. laevis* tadpoles occur in low-flow situations. It has been mentioned in Appendix A1 that if semi-static exposure is used, the concentration of test chemicals should be reported and that a 24hr renewal interval is ideal. The question is how practical and cost effective this is to measure the chemical concentrations in the water samples. Did the authors mean that in a preliminary study the dynamics of the mother compound in the water column must first be established?

Exposure procedures (control chemicals):
The experimental design seems adequate. However in a flow-through system the question arises regarding the effluent produced and how it should be handled/discarded. Although several types of flow-through

systems are potentially available, the authors don't give enough information on the diluter and flow-through system they used. They also don't describe and discuss the options of different semi-static systems that may be used or have been used by the German laboratory. Detailed SOP for using these systems are lacking and a laboratory that hope to do the AMA will find them in a vacuum.

Observe and Measure Prescribed Endpoints

David Crews: Pictoral references for histology readings, morphological measurements and image set up in AMA Test Method Appendices allow adequate standardization of measurements among multiple laboratories.

Catherine Propper:
1. Why is 10% chosen as an acceptable mortality rate when in the Phase Trials, 5% was the maximum acceptable mortality rate? No justification is given for this shift or for the 10% rate within the explanation of the test methodology.

2. Under determination of Biological Endpoints AMA Attachment 1 Test Method, beginning on page 8:

a. A URL link or reference to *Xenopus* staging with pictures should be provided within the test protocol.

b. Additional Observations (page 10):

ii. Behavioral Observations: If behavioral parameters such as uncoordinated swimming, hyperventilation, quiescence, etc are to be "observed," they need to be done so in a coordinated and quantifiable fashion. One methodology would be to do a 1 min focal animal observation on 3 animals/tank/ at day 7, 14, and 21. In the current protocol, there is no standardized way for making these observations and analyzing the results.

ii. Grossly Visiable Malformations: A list with pictures, if possible, of the usual gross morphological problems needs to be included in the protocol (kinked tails, bent backs, extra limbs). These problematic gross morphological outcomes should be included in the final evaluation at 7 days and 21 days and should have their own column in the data spreadsheets.

c. Under Test Initiation and Conduct: Day 7 (page 10):

If thyroid histology is to be conducted on Day 7, then it needs to be clearly stated here. If not, then there still needs to be a statement saying this subsample of the animals are to be stored individually in Davidson's Fix and then 10% NBF.

d. Under Data Collection and Reporting (page 12): Overall, the data tables supplied are adequate, but some additions, especially in summary tables would be helpful. Also, supplying a Quality Assurance Plan is necessary. It is referred to here, but not provided in any documentation.

Under Chemical Observations and data (page 12): Details need to be provided for how to collect the water for these determinations. Instrumentation should be identified, and SOPs provided as an appendix for everything except temperature and pH. This protocol should facilitate the ability of contracted and non-contracted labs to conduct an assay as similar to each other as possible. Also, if actual measures of test chemicals in the water are to be taken, then why might stocks also need to be measured? The way the protocol is worded now is very vague (states, "may be required") about whether the stocks need to be tested.

3. Attachment A2: The title needs to change so that the morphometric measurements come into play in the first part of the title. The title as reads emphasizes the histopathology. One suggestion is "Guidance Document on AMA Endpoint Sampling Part One: Technical Guidance for Morphological Sampling and Histological Preparation."

4. Attachment A2: Trimming of tissues. More detail is needed for how to remove the mandible for histological preparation. No detail is provided here.

5. Attachment A2: Image analysis. For each parameter that is digitally quantified, at least 2 measures should be taken and then averaged as there is some variation in how the lines are drawn. Also, one person should conduct all the measures across all treatment groups and should probably be blind to the treatment when conducting the measures.

6. Attachment A3: Some of the measurements of thyroid gland histology could be done via direct image analysis and direct quantification rather than semi-quantitatively by grade. However, this process is laborious and time consuming. The grading scheme, with proper training, and good preparation appears to be justified, and appeared to work for the assay in the validation data presented.

7. A Quality Assurance Plan document is mentioned, but none was provided. It should be an attachment or appendix with the AMA Test Method.

Hannes van Wyk: Developmental stage:-I am not convinced that all labs will extract the same criteria from N&F (1956) to determine the stage. Some guideline must be given and I feel detailed description of criteria used to stage a tadpole is necessary. The N&F (1956) document is not very friendly to read. How will one handle asynchronous development, making it difficult to stage a particular tadpole, using the standard trait set? Mention has been made of differential characters, advanced characters in the head region and arrested characters in the hind limbs. What set of characters are practical/important? The authors state that the staging is simple and clear-cut. I am not convinced about this.
Hind Limb length:- Gene expression studies show that the measurement of hind limb as endpoint make good sense. However, I am worried about not enough detail given as a SOP to measure hind limb in a standardized way. Especially when the limbs are long and well-developed, the line one takes when measuring may influence the outcome. Figure 1 in Appendix 1A is rather simplistic and does not show the real situation. The revised photo presented in the histological appendix does not solve this problem. Detailed landmarks are necessary for consistent results when applying a general bio-assay.
Body length and Weight:- In practice the opening of the vent is quite a difficult point to measure as well. Should one not use the base of the vent as an alternative measuring point/landmark?
Thyroid Gland Histology:- Numbers collected at day 7 and again at day 21 for histology will generate a large number of histological samples that need to be processed and eventually evaluated. The selection of individuals? 1) Why a day 7 sample? 2) The selection of samples seems rather complicated. 3) difficult to see it being practical to select randomly but also to try and stage match (later I see they actually recommended stage matching (see below). This will only be possible if one chemical is done at a time (with dilution replicates) and compared to a control. The absence of stages in treatment groups make stage matched comparisons difficult and will an extended control sample be necessary to generate same-batch control stages (see suggestion below).

Compile and Prepare Data for Statistical Analyses

David Crews: Please see previous comments on statistical analysis for PHASE 1 and PHASE 3 in section 2 of this review.

Catherine Propper: Under Statistical Analysis:
1. Mortality data cannot be analyzed by an Anova. Some form of G-test or Chi-squared will need to be employed.

2. A Mann-Whitney is employed if there are two treatments. For more than two treatments, first a Kruskal-Wallis should be employed first, followed by Mann-Whitney. Also, there is confusion across documents about how to conduct the statistics if the treatment effects are a linear verses non-linear dose response. Since in these types of studies (at least at low dose exposure) non-linear effects are often found, what type of statistic should be employed?

3. There is some confusion in the Phase trials about whether HLL should be standardized by body length or not. In the final data tables in the AMA Draft Test Methods, there is no mention of whether or not the HLL should be standardized. It needs to be made clear, prior to doing statistics, about whether this parameter should be standardized for final analysis and interpretation or not.

4. The Fig. 8.1 flow chart in the ISR has thyroid histology following only negative results in other areas. However, in reality, each of the tests conducted histopathology. If the EPA wants histopathology conducted always, then this assay needs to be placed higher up on the logical flow for data interpretation chart.

Hannes van Wyk: Not clear enough. Maybe the use of a diagramme will aid the understanding of the data grouping and analyses.

Report Results

David Crews: Performance criteria described in Table 4 of AMA Test Method (**File name:** AMA_Test_Method, pg. 14) provided detailed requirements of reportable data. Concern of small sample size at d21 compounded with mortality at this stage, as well as varying developmental stages within one treatment tank should be addressed. This reviewer did not evaluate the alternate static renewal design.

Catherine Propper: 1. Many data sheets are provided, but no guidelines for data interpretation are provided. In the three phase trails used in the validations, there were decisions made regarding the outcome interpretations, but in the test method, there are no guidelines. Once the data are collected and analyzed, how will they be interpreted? Summary tables like the ones used in Tables 4.5, 4.8, 4.12, etc. in the ISR should be provided to facilitate overall interpretation of the data. In these tables, instead of individual labs being columns, the dose of the compound used could be used across the top of the table. In fact, in the phased validation studies, one of the criticisms I have is that there was no information in the data summary tables of the doses considered to have effects. Again, this problem can be addressed in a final summary table provided in the AMA Test Method that allows for data interpretation across doses. Such a reporting system will also help in interpreting the data if non-linear effects are seen (see comments above).

2. Throughout the phase trials and in the Draft Methodology, there is no mention of how final decisions are to be made regarding the outcome of the test (mentioned also elsewhere in this review). It may be possible to combine all parameters measure and to apply a principle component analysis to determine the outcome of the exposure. Alternatively, consistent approaches to the data interpretation can be developed, and followed carefully. No matter the approach, it needs to be carefully outlined in the final Test Method.

Hannes van Wyk: The outcome seems clear. But, to come to a conclusion will take some interpretation, especially if the correlation between histological data and morphological data is weak.

Separating non-thyroidal toxicity from thyroidal effects will be problematic and criteria used vary vague. In particular, at the lower-end of agonistic and antagonistic effects.

Interpretation of Histopathology will have to be done by an experienced pathologist. Will it be possible to build this capacity in the laboratory or will expert scientists be contracted to do this part? How many amphibian pathologists do we have in the world, or will a human or wildlife pathologist equipped to do the screening?

More specific guidelines should be giver regarding the presentation of data. Can the reporting layout be made standard to ensure reporting of the data as well as assay performance data?

Please also make suggestions or recommendations for test method improvement.

David Crews:
a. Static-renewal. The alternative method, static-renewal, is described for insoluble compounds and high concentrations relative to the limits of water solubility (File name: Battelle_multi-chem_report) but is not used in subsequent Phase studies as the chemicals tested were water soluble. However, it is not described and so cannot be evaluated. If static renewal refers to the regular (periodic) replacement of the water in the tank, this is fraught with difficulties, not the least of which is the buildup of metabolic byproducts that can affect the endocrinology of the animals. Finally, if the Phase 1-3 testing was conducted using flow-through systems, alternatives such as static renewal should be disallowed until comparable tests for intra- and inter-laboratory QA/QC are conducted.

b. Source of animals and selection of spawns. The source of the adult animals (pg. 5), and the "best spawns" (pg. 5) are a concern. That is, it appears that all of the egg masses will be collected together and a selection is made. This could result in only a few of the mating pairs producing most of the tadpoles used in any specific study. This may be mitigated by the treatment of the selected spawns being treated with a 2% L-cystein solution and then combining the larvae, but it would be preferable to use ALL spawns produced treat them all, and selecting the larvae after they are freed from the jelly coat. The large discrepancy in the animals in the control groups in Phase 2 illustrates the importance of the source of animals.

c. Analysis of food. The quality control (QC) of the food offered to the larvae/tadpoles are not described and are of concern. Is there documentation and analytic verification available for each production? While the same vendor is being used (Sera GmbH), it is well known that batches of commercially available foods for laboratory animals can vary significantly. Further, if the food is produced by multiple facilities of the same company, and thus purchased by different testing facilities, this can be a significant source of variation between testing facilities.

d. Maximum Test Concentration. The highest test concentration, or MTC, is defined as the highest test concentration of the chemical that results in less than 10% acute mortality (pg. 7). This is a concern. It is stated that if prior empirical acute mortality data are not available or sufficient information is not available to develop regression models to estimate the MTC, then a 96 hr LC50 test will be conducted. The LC50 traditionally is defined as the lowest concentration that results in 50% mortality, but it is not clear if this is how it is defined here. If, however, this is the definition used here, then the MTC would be calculated as being 1/3 of the LC50. The lower concentrations to be tested would be calculated as a dose separation of 0.33-0.5 (max-min). This does not correspond to best practice NOAEL calculations.

e. Dilutions: Given the concern about low dosage effects, it is not clear why the AMA advises that only three dosages of the test chemical be used. This is particularly puzzling when in the Phase studies four or more dosages, spanning a full log unit or more, were used.

f. Initial sample. For a d0 measure, approximately 20 individuals will be measured for WBL. It is not clear if these 20 individuals will be reintroduced into the test population for distribution into the tanks, or whether they will be used to obtain the other stated measurements (see above).

g. Sample size. A sample of 5 tadpoles will be taken from each tank on d7, for a total sample size of 20 tadpoles for each treatment/dose. This allows for up 15 remaining tadpoles per tank for a second terminal sample on d21 (or a total of 60 tadpoles for each treatment/dose if there is no death or disability). This is unlikely to be the case, and the issue of how the requisite 20 individuals will be selected for in-depth analysis for the d21 sample is considered. If size matched samples are to be used as stipulated, why was the most advanced stage selected for analysis? Also, why is this same criterion not applied to the d7 sample as a distribution of stages are likely to be present as well (although perhaps not as wide a range)?

h. Asymmetrical limbs. The body plan of most animals, including frogs, is not symmetrical. Although differences can be slight, they are present and have been shown in various studies to be important mechanistically as well as evolutionarily. One side should be selected and it be mandated that this side only be measured.

i. Scoring of slides. The Phase I and Phase 2 studies addressed the issue of intra- and inter-laboratory variability. Although the results of both sets of studies indicated this variation to be minimal, with the "response profiles of the various endpoints were different for the individual test substances but reproducible across laboratories", this reviewer still has a concern that any initial screen be conducted in a non-blinded fashion, this must be limited to evaluation of the quality of the sections and their suitability for measurement. It is mandatory that "any potential compound-related findings will be re-evaluated by the pathologist in a blinded manner prior to reporting such findings" (pg. 6). The following terminology "when appropriate" is absolutely inappropriate. The caveat that "Certain diagnostic criteria, such as thyroid gland hypertrophy or atrophy, cannot be read in a blinded manner due to the diagnostic dependence on control thyroid glands" (pg. 6) can be mitigated by having a set of standard slides that are distributed to all potential contractors. The images provided in this document are excellent and could serve this purpose at least initially. Finally, there is a need to assess inter-observer reliability both within the same laboratory as well as across contract laboratories. There should be separate Quality Assurance/Quality Control (QA/QC) performance guidelines.

j. Radioimmunoassay. As stated by DeVito et al. (1999) above, a less costly and time-consuming alternative is available. In these instances, whole bodies or heads can be extracted and TH concentrations can be assayed using either radiometric or ELISA methods.

J. David Furlow:

a. Quantitative PCR for gene expression markers of thyroid hormone action (such as well characterized, broadly expressed TH response genes like TR□ and TH/bZIP, or markers of disruption of the HPT axis like TSH□ or NIS) would provide a highly quantitative assay that allows the investigator to assess proper thyroid hormone signaling in specific tissues. This aspect of the metamorphosis system is arguably as well, if not better, developed than for estrogen or androgen action in rodents. Futhermore, newer transgenic models are being developed that provide fluorescent or bioluminescent markers of thyroid hormone action in Xenopus.

(For example: the system being developed by Barbara Demeneix's group and the start-up Watchfrog in France: Turque N, Palmier K, Le Mével S, Alliot C, Demeneix BA. A rapid, physiologic protocol for testing transcriptional effects of thyroid-disrupting agents in premetamorphic Xenopus tadpoles. Environ Health Perspect. 2005 113(11):1588-93; Fini JB, Le Mevel S, Turque N, Palmier K, Zalko D, Cravedi JP, Demeneix BA. An in vivo multiwell-based fluorescent screen for monitoring vertebrate thyroid hormone disruption. Environ Sci Technol. 2007 41(16):5908-14.).

b. Since tail resorption is not an endpoint of the whole animal based assay, tail organ cultures are well established, highly reproducible and quantitative, and dose responsive, and would serve to detect interference of compounds directly at a target tissue.

(For example: Schriks M, Zvinavashe E, Furlow JD, Murk AJ. Disruption of thyroid hormone-mediated Xenopus laevis tadpole tail tip regression by hexabromocyclododecane (HBCD) and 2,2',3,3',4,4',5,5',6-nona brominated diphenyl ether (BDE206) Chemosphere. 2006 65(10):1904-8; Ji L, Domanski D, Skirrow RC, Helbing CC. Genistein prevents thyroid hormone-dependent tail regression of Rana catesbeiana tadpoles by targetting protein kinase C and thyroid hormone receptor alpha. Dev Dyn. 2007 236(3):777-90; Furlow JD, Yang HY, Hsu M, Lim W, Ermio DJ, Chiellini G, Scanlan TS. Induction of larval tissue resorption in Xenopus laevis tadpoles by the thyroid hormone receptor agonist GC-1. J Biol Chem. 2004 279(25):26555-62; Lim W, Nguyen NH, Yang HY, Scanlan TS, Furlow JD. A thyroid hormone antagonist that inhibits thyroid hormone action in vivo. J Biol Chem. 2002 Sep 20;277(38):35664-70.)

Catherine Propper: I have incorporated most of my suggestions in my comments above, and I will summarize the main points below under section 8.

As asynchronous development is an important endpoint as pointed out a least twice in the summary documents, it will be critical to provide labs with a clear-cut standard operating procedure for scoring this issue and analyzing and interpreting the data.

Hannes van Wyk: More detailed SOPs are needed. The earlier suggestion by the German, French and Irish scientists (DPR) that a short-term gene expression study be included seems to make sense. The initial response that the level of complexity of this technique and the difficult interpretation of the data may not be valid since histological interpretation seems to be rather complex as well, although cheaper to produce. Investigating laboratories could out-source these aspects to specialists (will probably have to do it in the case of histopathology anyway).

Discussion of refinements suggested in the ISR:-
Dietary regime:- I agree that the feeding regime must be standardized or monitored in relation to a few growth performance checks, say at 7 days and 14 days in the control groups.
Water iodine levels:- I agree with this suggestion. The question is, has this problem been researched adequately?
Dose levels:-I agree with this suggestion (see below)
Stage matching for histopathology:- I agree with this refinement, however, the comparison is with the Control group and in some cases you will not have matching stages (either in antagonist or in agonist groups). Two possibilities may solve this problem: 1) if the performance criterium of Control stages being around NF 57 then comparisons could be made to known histology documented from all developmental stages (Atlas approach), but, if it is better to compare with internal control, then initial control sample (groups should be increased and representative stages sampled at certain times to facilitate stage matching with controls. Following 21 days, remaining Controls can be maintained to reach stages reached in agonist groups. At least for this batch stage matching will be possible. This problem only occurs when working with strong antagonistic and agonistic chemicals.

Improved Data interpretations:- Agree with the suggestion, but would like to know why only use "advanced development" to get a "Yes" and therefore exclude the need to do histology? Why not also include "advanced inhibition/retardation" to get a"Yes"? I know there was a concern that the histopathological assessment is time- and specialist-consuming and should therefore be limited. But, without direct evidence of some kind, for example, molecular (thyroid receptor (TR) expression)) or histological evidence the risk of getting a false positive (agonistic or antagonistic) seems to be greater? Refer to Table 8-3 for T4. If I understand correctly all labs will conclude "active" after noting advanced development. However, the histology only supported this conclusion in two of the labs. I can see that the compound will still move to Tier 2 but at least more direct knowledge will be available regarding the histopathological picture.

Another suggestion: I am of the opinion that collecting of material (in RNAlater for example) during the exposure phase (either independently at 48 hours or after 7 days) could add another level to Figure 8-1. QPCR technology is becoming more and more routine now and could greatly aid as a last step just to make sure you don't have false negatives.

Comment on the Strengths and/or Limitations of the Assay in the Context of a Potential Battery of Assays to Determine Interaction with the Endocrine System

David Crews: What follows refers only to the primary document, Test Method for the Amphibian Metamorphosis Assay (File names**:** AMA_Test_Method; AMA_Test_Method_Appendix_1; AMA_Test_Method_Appendix_2)

Strengths of the assay

a. It is commendable a flow-through method is recommended. This avoids the problem of buildup of metabolic byproducts that can influence the stated endpoints as may occur in the static-renewal method.

b. The use of widely accepted developmental staging for *Xenopus* development.

c. The use of a defined time window for exposure.

d. The use of computer-assisted software for microscopic determinations.

e. The use of standardized histological protocol.

f. The use of standard histological slides to facilitate evaluation of thyroid histology.

g. The issue of sample selection for the terminal sample (d21) is considered and detailed.

h. The issue of variation within and across laboratories has been addressed in rigorous manner.

i. The statistical evaluation and power analysis as guiding principles for implementation of the AMA is excellent (File name: Power_Analysis).

Limitations of the assay.

a. Low Dose. Recommend dosages spanning at least one full log unit and having at least four concentrations to determine true nature of the dose-response.

b. Threshold. This protocol does not allow for this important determination.

c. Mixtures. This protocol does not allow for this important determination. "Recent findings of a rather strong activity of BP-2 in *in vitro* assays and the marked difference in the severity of BP-2 effects on the thyroid system in two different laboratories could be interpreted that the actual potency of BP-2 to disrupt thyroid system function is strongly dependent on iodide availability." (pg. 70) (File name: OECD_Phase_3_Draft_Report) It is possible that the potency of other chemicals may depend on differential iodide concentrations.

d. Mortality vs. Evolutionary Death. The present EDSP focuses on the individual in its own lifetime. This is valuable information, but says little about the impact of the chemical on the population through time (proximate or ultimate). One measure is whether an individual will breed. If the individual does breed, but its young do not develop properly and do not breed, than the overall result in terms of the population is the same. If the goal is to have a means of evaluating the impact on compounds that have an impact on thyroid function for wildlife and human health, then it is the latter issue that is pertinent.

e. Sex Differences in Sensitivity. If one goes to PubMed and inputs "sex differences in thyroid function", 132 citations come up in the primary literature. If this is further refined to "sex differences in thyroid function, development" 15 papers are cited. Typical is that of Ng et al., (2007) findings that female infants with thyroid ectopia have significantly higher thyroid stimulating hormone (**TSH**) concentrations than do males and significantly lower circulating concentrations of plasma T4 were significantly lower than in males. Since the animal is being sectioned for histology, it would be a simple (but adding to expense) addition to look at the gonads. Given that the tested compounds may also influence differentiation of the gonads (see below), it would also be necessary to use standard genetic markers for sexing the tadpoles (see.

f. Multiple Target Organs. Hyperthyroidism induced by PTU or methimazole, also acts on developing gonad, specifically in males on the Sertoli cells. It has been known since 1925 (Rickey) that thyroidectomy eliminates sexual activity in male rats and in more modern experiments in both mammals and fish the Sertoli cells early in testicular differentiation have abundant receptors that decline markedly after sexual maturation (Cooke, 1996; Kirby et al., 1992; Matta et al., 2002; Schultz et al., 2005). Thus, the observation in Phase 3 that E2 caused male-to-female sex reversal without affecting other measures including the histopathology of the thyroid should be considered seriously.

g. Procedure for Training of Pathologists. Need to assess inter-observer reliability both within the same laboratory as well as across contract laboratories. Best course would be to require that a standard set of slides/images be provided to each contract laboratory and Quality Assurance/Quality Control guidelines be developed and adhered to with no exceptions.

h. Measuring the Same Side. The animal body is asymmetrical and so it would be necessary that the same limb be measured on each tadpole.

i. Sample Sizes. The issue of the selection of tadpoles for the sample dates (d0, d7, and d21) are considered above. Here though I raise another issue. If size matched samples are to be used as stipulated, why was the most advanced stage selected for analysis? Also, why is this same criterion not applied to the d7 sample as a distribution of stages are likely to be present as well (although perhaps not as wide a range)? This is unlikely to be the case, but raises the issue of how the requisite 20 individuals will be selected for in-depth analysis for the d21 sample. Further, what is to happen if mortality and disabilities may be such that adequate animal numbers will be available to obtain a meaningful sample?

j. Standardization of Food vs. Potential EDC Content in Food. Specify parameters of the food by analytic chemical analysis and make each contract laboratory supply documentation of having met these criteria with each report.

J. David Furlow:
Strengths:

a. The assay uses an intact animal model that is highly sensitive to thyroid hormone rather than relying solely on cell lines or biochemical assays to predict effects on animal physiology.
b. Chemical analyses are required to make sure compounds meet nominal values.
c. Careful analysis and maintenance of water quality conditions are described to eliminate non-specific effects on metamorphosis.
d. At least one well documented for histopathological assessment is included for comparison to external morphological changes.
e. Two important issues poorly covered by the Agency to date in toxicity evaluations are both addressed here: thyroid hormone synthesis/action and amphibian biology.

Limitations:

a. There is no (or limited) mechanistic component to the assay. It would not be difficult to incorporated gene expression analyses and hormone measurements to the assay.
In Table 1-1 of the ISR, thyroid hormone receptor binding assays and transcriptional activation assays are not listed as additional tests whereas androgen and estrogen receptor based assays are listed as planned.

b. The animals only develop up to a stage just prior to climax; therefore only acceleration or inhibition of hindlimb growth make up the bulk of the analytical component of the assay not effects on tail resorption which may be more sensitive to perturbations in TH levels.

c. Mixtures effects are not accounted for at all. This issue is something that the EPA should start addressing sooner, rather than later. Is the assay robust enough to detect a reversal of T4 effects by IOP, for example?

d. The effects of selective hormone receptor modulators (eg tamoxifen) can be tissue and even species selective in their actions, and endocrine disrupting chemicalss may well follow suit. In this assay, essentially all of the analysis is focused on the hindlimbs (due to the nature of the developmental staging criteria and the direct hindlimb measurements) and thyroid gland histology. Selective effects in specific tissues could be readily determined by incorporating gene expression analysis.

e. Finally, the suitability of *Xenopus laevis* as a surrogate for other amphibians may be questioned. *Xenopus laevis* is a primitive amphibian that does not have a fully terrestrial adult stage, and is not native to North America. In addition, many studies have differing strains of rats can show wide differences in responses to endocrine disrupting compounds and there is essentially no data to my knowledge about this issue in amphibians.

Catherine Propper:
a. The strength of this assay is in its ability to determine whole animal disruption of thyroid hormone-related physiology. One weakness is that the assay itself will not determine how the disruption is occurring.

b. One limitation of the assay is that animals are not dosed throughout development. Such testing may lead to increased sensitivity of the assay (see comments above).

c. A major limit to this method is the number and choice of doses used in the assay. Little justification for the dosing decision-making process is given. The doses are decided based on the overtly toxic dose. The ISR needs to present a clear rationale for this dosing approach, and it needs to be made in light of the literature in the field. The decision to only go with three potentially very high doses that do not even differ by even 10 fold is a mistake. First, the literature in endocrine disruption demonstrates time and again that there are non-linear responses, especially at very low doses. Second, environmental exposures to many chemicals in the environment are occurring at the part-per-billion or even part-per-trillion levels. The current dosing regime for this assay would most likely be well above these levels. Last, there is the issue of non-linearity of response, especially at the lower doses that are important given the risk of exposure to human and wildlife populations are mostly at low doses. In summary due to the non-linearity of some dose responses and the fact that a very low dose can have more impacts on endpoints than higher doses, these doses need to be evaluated. The AMA should be sensitive enough to pick up on these low dose and non-linear responses.

d. Because of the issue of non-linearity, this methodology needs further development with how to deal with no- linear dose responses. The report was unable to really respond to the occasional non linear response, yet in many endocrine disruption studies, the finding of non-linearity is the case. A clear approach is necessary. For instance the final scientific review panel may state that if any dose has an effect, the result is a positive. Alternatively, they could decide that if two of the 3-4 doses tested need to be positive before they determine an effect. How will these types of non-linear results be interpreted?

e. The methodology (and in fact the validation trials) do not provide much information for reporting the dose effects. The overall reporting is a yes or no outcome in the reporting tables for the phase trials with no information provided about the lowest effective dose level. Dose effects need to be taken into account in the final reporting for the assay.

f. One last limitation is the lack of how this assay can address the issue of exposure to complex mixes. The field of ecotoxicology is still in its infancy regarding evaluation of the complex mixes which are what all organisms are really exposed to. Furthermore, mixes can interact with each other to lead to endpoints that individual compounds will not. Even thyroid hormone and estradiol interact (see comments elsewhere in this peer-review). Can this AMA protocol be applied to testing for understanding the thyroid hormone disrupting capacity of complex mixes? Even if the EDSP purpose is not to test mixes, others in the field will want to adapt these protocols as closely as possible to their studies.

Hannes van Wyk: I agree with the discussion on potential limitations listed, but also underline that several of these represent knowledge gaps. The use of non-mammalian models as early warning systems to human health still has to go a long way. However, the appreciation of interaction between environment and organism will flow from such aquatic non-mammalian models. While it is true (point 2) that morphological and/or molecular responses may be different in developing young and adults, the effects at the developmental level by several EDCs are the most dramatic, both at short-term and long-term levels. Surely, potential endocrine disruption result in concerns at both levels? Regarding point 6, I am a bit concerned that we the level of knowledge regarding the normal histological profile of the developing tadpole along with the tissue specific gene expression profiles are generally lacking and therefore represent a major gap.

Another concern is the fact that we start the breeding by using high doses of hCG. Do we really understand the consequences of these doses for the mother (thyroid system) and the changes in aspects of maternal transfer, therefore impacting on the developing tadpole? This screening tool compare against a control, but maternal transfer may affect response sensitivities towards unknown compounds (false positives?).

Richard Wassersug: The greatest strength of the assay come from the amount of work that the EPA, its partners and its contractors have put into developing the assay over the last decade. They have made major progress in developing a reliable amphibian metamorphosis assay. Given the concerns about endocrine disruptors in the environment, this effort was appropriate. There are, however, some holes in the protocol about how to perform the assay. As stated extensively above, important variables in the execution of the assay are missing from the documents provided. The biological relevance has to be qualified given how different *Xenopus* is than all of the non-pipid anurans in the world (see #3 above).

Provide Comments on the Impacts of the Choice of a) Test Substances, b) Analytical Methods, and c)Statistical Methods in Terms of Demonstrating the Performance of the Assay

David Crews: The choice of test substances and methods were reasonable.

J. David Furlow: The choice of test compounds is highly appropriate, aside from the previously mentioned limitation on the ability of the assay to detect mixture effects. In the interlaboratory exercise in particular, the choice of perchlorate, T4, and iopanoic acid covers three distinct mechanisms of action is highly appropriate.

Catherine Propper:
a. The choice of the tadpole metamorphosis system as a test assay is outstanding given the knowledge base of the system, and the relative ease of use and data interpretation. The comments below are to the details of the method and not to the overall utility of the assay. Once the methods are standardized and clearly detailed, this assay will undoubtedly provide a useful measure for thyroid hormone disruption.

b. In the development of the assays, one lab used static renewal methodology while the others used flow through systems. Ultimately, the ISR states and the AMA Test Method Draft recommends the flow through system with little or no justification based on the studies. The data clearly demonstrate no difference between the two systems in control performance. Also, in the validation no data are provided for the actual dose received in the static renewal system. If static renewal is to be allowed, it is critical to know if the concentrations of chemical treatment (dose the animals receive) are equivalent between the two systems.

c. As mentioned elsewhere in the review, the assay is validated using compounds that are known agonists or antagonists of thyroid physiology. Also, a presumed non-thyroid disruptor was evaluated (estradiol; see problems with data interpretation above) along with a weak distruptor (BP-2) at fairly high doses. It would be useful to have one more validation step using a pesticide of some form that has known thyroid hormone disrupting effects at environmentally relevant levels.

d. The interpretation of results with a compound like IOP is very interesting, and needs to be carefully evaluated, as some of the compounds likely to be tested via the EDSP may have such complicated modes of action. The results on HLL may be difficult to interpret given the impacts of such compounds also on body length, but then it is possible to do the analysis as an index: HLL/BL.

Hannes van Wyk: A concern to me was that during this validation testing only limited potential mode-of-action modulation was tested. More attention should be give towards selecting controls representing different mode of actions, especially in a complex system like the thyroid. By including IOP in the Phase II series showed that at the developmental level unexpected results can be found. For this assay we need causal relationships between morphological endpoints and different moed of actions. Moreover, I feel

strongly that the link to possible use of the AMA in screening mixtures, and environmental samples at the Tier I level examples must be made. To what extend could the AMA be used to screen these complex samples?

The range of both agonistic and antagonistic representatives operating at different input sites (different modes of action) was rather limited and questions remain.

Richard Wassersug: The tests subjects used to demonstrate the performance of the assay were appropriate as were most of the methods used. However there are still some methodological problems, which are discussed extensively above.

Provide Comments on Repeatability and Reproducibility of the Results Obtained with the Assay, Considering the Variability Inherent in the Biological and Chemical Test Methods

David Crews: This is a major flaw of the material provided and is detailed in the above comments.

J. David Furlow: One of the major concerns about the assay is the degree of inter-laboratory consistency. The first concern, regarding the variability in the progression through metamorphosis by the controls, appears adequately addressed by the lower amount of feed provided in Laboratory 3 (seen in Tables 5 and 6, Interlaboratory report) . Aside from the general delay in metamorphosis and high degree of variance seen in animals from that laboratory, the degree of consistency within a given laboratory is in fact quite good and the investigators are to be commended.

The second concern is that while overall trends are observed (ie T4 accelerates, perchlorate and IOP delay), there is surprising inconsistency among the laboratories. For example, only labs 3 and 4 detected a significant effect on hindlimb growth and developmental stage at the two highest levels of perchorate tested by day 21 whereas three other labs did not (Table 15, p. 46, Interlaboratory report). Since laboratory 4 apparently had adequate control animal development this cannot simply be due to feeding differences. While the T4 experiments were more in agreement, in the IOP studies, Laboratory 5 shows no effect at all of IOP at either day 7 or day 21 and Laboratory 3 shows a significant effect at 7 days with regard to hindlimb growth (Table 38 p. 70; Interlaboratory report). Furthermore, it was highly surprising that despite effects on hindlimb growth reported in all laboratories except laboratory 5, no significant effects on NF staging were reported. (Table 37 p69 Interlaboratory report). Also, the progression of control animals through metamorphosis by 21 days was remarkably different in this study (Lab 1 ~58, Lab 2 ~59, Lab 3 60-65, Lab 4 ~59, Lab 5 60-62).

Finally, the summary of the thyroid histopathology results are somewhat confusing. In the ISR, p. 60, Figure 5-1, 100% of all glands from all animals where scored as having follicular cell hyperplasia in laboratory 3 whereas the other laboratories scored a generally increasing dose responsive effect. Indeed, across treatment groups, there is a trend of high incidence of abnormality by laboratory 3. Does this reflect lack of experience of the pathologist with scoring amphibian thyroid glands or in the growth and treatment regimen? It might be useful to have the slides from laboratory 3 scored by pathologists from other laboratories, or to have used one pathologist. This concern is amplified in the T4 responses where there is even greater inconsistency between laboratories.

Based on these observations, the consistency of findings across laboratories remains a major concern for the future viability of the assay system.

Catherine Propper: Overall, the interlab variability was minimal, however, there was some variation and testing may need to be conducted independently in at least two separate labs.

Hannes van Wyk: Repeatability and reproducibility of the results obtained with the assay, considering the variability inherent in the biological and chemical test methods
OECD Phase I study:- The repeatability of the AMA among three different laboratories showed some consistency. The outcome of the Phase I study corresponded to predicted results, especially in the higher concentration exposure groups. In this comparative study the fact that similar results were obtained in spite of variation in protocols used, show / confirm that the *Xenopus laevis* metamorphosis assay (XEMA) is a robust assay. Control compounds affected the thyroid histology as predicted. PTU exposure showed that chemicals affecting the iodine transport system will noticeably inhibit the TH output and thereby affect the functional aspects of the thyroid. Moreover, TH will result in increased thyroid activity. Compared to the Control tadpoles, significant inhibition and stimulation occurred. Differences in protocols used among the three participating laboratories clearly suggest that in spite of these differences, comparable results could be generated.
Multi-chemical study (USEPA):- The outcome of the known control chemicals (T4 and PTU) supported the results of the Phase 1 study. The results therefore confirm repeatability using the AMA protocol. In this study it was difficult to find the motivation (reason for inclusion) of most of the other chemicals (Discussion of Appendix H). The results of this study show that in many cases the understanding of the mode of action of lesser known chemical affecting the thyroid axis will need multiple endpoint studies. It is clear that the AMA represents a starting point only. In this study the importance of using the histopathology endpoints was underlined. One aspect that worried me is the chemical application (for example PCN). It seems that although theoretically sound it may be difficult for consulting laboratories to do exposure studies. Too little information is given on this aspect. I am still not convinced that body weight is a good indicator of thyroid effects. In this study histological observations were valuable and it shows that a combination of endpoints must be used. However, the studies present histopathology results as descriptions and it is difficult for the reader to visualize disruption. The question remains, how practical will it be for a reasonable inexperienced research team to evaluate histo-pathological endpoints? Another concern is that it was not clear whether the 21 day exposure starting with day 51 tadpoles were the better option (especially when evaluating histo-pathology). The compensatory hypothesis surfaced and more research is needed on this aspect. I am a bit worried that the limited number of endpoints, and the mode of action associated with these endpoints were not adequate to show thyroid disruption in some of the selected chemicals.
Inter-laboratory study (Phase II):- In this study all the knowledge and experience gained from the previous studies were used to standardize protocols. To solve some of the reproducibility issues more detailed descriptions of protocols were used. However, I think it is still not well-described and would lead to measurement errors and increased variation. A second concern that was highlighted was the variation that occurred in the development of control animals in one laboratory. The report attributes this variation to differences in feeding regimes. Added to this is the staging at of stage 51 tadpoles; could it be that inaccurate stage determination (stage 51) result in different developmental stages as early as day 7 of the exposure? Although the Perchlorate control gave reasonable consistent results I was surprised by the inter-laboratory variation in results. I am not sure these variations were adequately addressed. One question that comes to mind is the aspect of observer error or reproducibility. Was the scoring of observers validated internally and between laboratories? The histological reading and scoring could be great source of variation. In general it seems that Perchlorate could be used as a standard control. How much regarding thyroid axis disruption can we read into general morphological endpoints like body size and weight? Just in control tanks we see so much variation in these growth parameters. The developmental endpoint in the Thyroid exposures corroborated previous studies and showed that this positive control worked well. However, the inconsistent results using the histological criteria were somewhat surprising. To what extend could this result be attributed to the fact that tadpoles were selected for histology on a random basis,

therefore potentially including different NF stage tadpoles in the sample? Are we assuming that the histological picture is independent of developmental stage in exposed groups? Stage matched comparisons would have help answering this question. In the IOP control exposure the asynchronous development showed that staging problems may arise with certain chemicals. The question is would the gene expression studies (short-term study proposed by German group at som point) not helped to explain some of these results. I just get the feeling if endpoints respond strange or not at all in a limited array of endpoints, so much are lost. In this case the histology did not respond clearly either. It seems that if the mode of action is largely unknown then unpredicted results will make interpretations difficult.

The conclusions of the **Phase I and II** studies seems valid and underline certain concerns mentioned earlier. One aspect that increases the work load is the inclusion of a day 7 sampling. It seems that in the agonistic exposure (T4) growth parameters showed some sensitivity and helped interpretations when later compensatory effects came into play. However, whether the histological investigations at this stage made a valuable contribution was not clear. In general day 7 data seems to help the researchers to make an early assessment of how the exposure is going and it seems that the sampling at this time could be limited to save on labor.

OECD Phase III:- The stated goal of this exposures was to establish whether AMA could effectively indicate whether a compound needs further testing at Tier 2 level. The selection of compounds, for example 17β-Estradiol was not well-motivated. The statement that it is a potent endocrine disruptor is very general. To me endocrine disruption point to a mode of action or specific functionality and to include E2 only because it is a potent estrogenic EDC does not really make any sense. I presume the goal was to screen chemical with low predictive thyroid activity, but high activity in other areas of endocrine disruption? Was E2 included as a control since there is some indication that Benzophenone (BP2) is estrogenic? In the BP2 exposure study it was concerning that the two labs gave different results. It was attributed to differences in iodide concentration in the water. This underlines the value of standardizing all aspects of exposure when doing an inter-laboratory study. It was not clear why the difference in dilution water?

Other published studies:-From the literature it seems that results of known control chemical corroborate the results of the inter-laboratory studies, although in most cases histological studies were excluded from these.

Overall-comparison and Conclusions: -I suppose most data suggest that when using certain control chemicals (T4, PTU) that the reproducibility of the AMA as a screening tool has been well demonstrated. This was especially true in the Phase I and II studies. Concerning was that not all aspects were always controlled for. Moreover, when conducting the inter-laboratory study using weak thyroid modulators, it seems that the consistency was lost.

The result of the inter-laboratory studies was the formulation of clear performance criteria. I agree it would reduce variability and ensure some form of assessment regarding performance of the metamorphosis assay. However, little attention was given to the source and time in captivity of the *Xenopus laevis* breeding pairs that a laboratory may use. Minimum median developmental stage of controls at the end of test may not be reached but the comparison between controls and experimental (unknowns) could still suggest further testing (Tier 2). The screening of the chemical is the main goal. Another question that should be asked: Is it necessary to include known agonist and antagonist controls? The implication is that the test laboratory always starts with three or four exposure groups. It seems that a laboratory could run these controls to determine capacity but that once this has been shown these could be exclude. The suggestion is that the performance criteria are applied after the 21 day trial. It seems from the studies conducted that one could include day 7 as some indicator? What about putting in a developmental check

in the Control group at 14 days as well? To run the test to its completion and then assess performance seems unrealistic.

Richard Wassersug: One of my greatest concerns in the AMA documentation is the high variance in reproducibility of the results obtained from the various labs during the various test phases. I am disquieted by the little attention given to the variance between the labs, when their protocols were (supposedly) identical.

Most of the chemicals used in these studies were well known inhibitors or accelerators of metamorphosis. The fact that inhibition and acceleration were seen in the test results is, of course, exactly what one expected. I did not expect, however, the variance in the reports between the different labs. It is bothersome that more effort was not made to explain the inter-laboratory variance.

Please comment on the overall utility of the assay as a screening tool, to be used by the EPA, to identify chemicals that have the potential to interact with the endocrine system sufficiently to warrant further testing.

David Crews: Before the AMA can be used as a screening tool that is open to contract laboratories, the issues raised above should be addressed. The bottom line is that the AMA is not suitable as a screening tool for endocrine disrupting compounds.

J. David Furlow: The assay as designed should be able to detect the presence of that, by themselves, can disrupt the normal progression of metamorphosis, and thus by inference, disruption of some point along the hypothalamic-pituitary-thyroid axis or thyroid hormone activity in peripheral tissues. It is an outstanding first step in developing a whole animal bioassay for thyroid hormone system disruption.

However, while there are many excellent aspects of the study design and presentation, several issues summarized above currently preclude the assay's use as a routine screening assay, most notably the high degree of interlaboratory variability, the lack of assessment of endpoints other than basically hindlimb development and thyroid gland appearance, and the recommended dosing regimen is too narrow to discriminate between general toxicity and specific endocrine disruption.

Catherine Propper: Overall, the AMA will be a useful screening tool for testing compounds and complex mixes for thyroid hormone disruption. There are some details that need to be added or clarified within the assay protocol itself, and some addition information/validation that might prove useful. These issues (all brought out above) are summarized below:

Summary Points:
1. The outcome of "Thyroid Active" needs to be divided into two categories to provide information regarding whether a compound has agonist-like or antagonist-like activity.
2. More detail is necessary in the set up of the assay and in the delivery of compounds.
3. Clear consistent control water from DI or e-pure water with salts and iodide added back must be used rather than region-specific tap water.
4. More detail is needed in the raising of tadpoles to stage 51, and in what type of water should be used for the first days of growth, and at what stage to switch to the water for culturing the animals during the study period.
5. Tadpoles from multiple spawns should be divided among all tanks and treatments.
6. There is a strong need for clear guidelines for data interpretation. The phase trials provide tables with a thyroid disruption +/- scheme, but the interpretation of the presented results across the three trials are not

consistent. For example, would the testing lab conclude that a compound is thyroid disrupting if at least 2 criteria are met? How about 3? In other words, how will those summary tables be used to determine whether a compound has thyroid-like activity, blocks thyroid hormone function, is not thyroid active or is toxic. Again, there was inconsistency in data interpretation across the Phase trials.
7. Dosing needs to be over a wider range and needs to have some treatments that are within predicted exposure levels for human/wildlife populations (low ppb range).
8. Mechanisms for reporting dose outcomes and overall dose limits of sensitivity need to be developed.
8. A final validation step needs to be undertaken to evaluate one or two more compounds known to impact thyroid hormone function. These studies should compare the outcomes of the doses determined as described in the AMA Draft Test Method to environmentally relevant doses.
9. Concern exists for the interactions of these compounds. One of the main limits of any of the EDSP assays is that they do not address the impacts of complex mixes of compounds. No organism is exposed to any one compound, and it needs to be noted in the final version of these assays that a negative finding for the potential for endocrine disruption cannot preclude that the compound might interact with others to have endocrine-relevant impacts.

Hannes van Wyk: Overall utility of the assay as a screening tool to identify chemical that have the potential to interact with the endocrine system.
As pointed out in the objectives, the AMA as an *in vivo* screening tool represents a multi-endpoint model system. This assay integrates effects. Its greatest drawback is the time factor. Most organizations or researchers interested to screen compounds for more definitive testing are focusing on rapid tests, receptor binding assays or specific biochemical elements in certain pathways. From this perspective, the AMA is a long and labor intensive (expensive) bio-assay at the Tier I level. Indeed one may reason that we are paying high costs for an extensive complex bioassay with endpoints that are reasonable difficult to assess (especially the histological endpoints. However, the simplicity associated with the aquatic exposure of developing *Xenopus laevis* tadpoles offers unique opportunities to screen environmental chemicals. In contrast to mammals, tadpoles are assessable throughout their development and differential gene expression profiles exists throughout the developmental programme, making the selection of specific exposure windows more simple and controlled. Although the use of *in viv*o models for Tier I screening has been criticized it gives a more integrated response system. Therefore, I am convinced that the AMA has great advantages in identifying chemical that interact with the thyroid system. In combination with specific molecular end-points confident assessments will be made that will greatly aid the sorting of potential EDCs. Advancing to the screening of mixtures and environmental samples should be rather simple. The AMA lies at the interface of rapid, very sensitive and very specific in vitro assays, but with the advantage of an integrated in vivo response system, closer to the true picture of endocrine modulation. In addition, the AMA continues to contribute to the understanding of the role of thyroid hormone in vertebrate development, including mammals and humans. Since *X. laevis* has been a classical model system in embryology studies for decades, and the fact that several aspects of its endocrine physiology is well-understood together with recent advances made in the molecular field (creating specific tools to understand developmental stage-specific responses to TH and EDCs) the utility of the AMA assay is valuable and will allow for making links to more detailed studies regarding endocrine disruption.

In conclusion, I am of the opinion that the development and validation of the AMA using *X. laevis* as model has come a long way and should be implemented. However, it should be remembered that it is a qualitative screen. The refinements suggested should be incorporated and acknowledged that future refinements will continuously arrive to be incorporated. The AMA is a valuable and unique opportunity to use a rather simple *in vivo* system at the Tier I level.

Richard Wassersug: Despite all the concerns stated above, I feel that the EPA should accept the AMA—*with expansion of its protocol documentation*—as a screening tool for chemicals that may have the

potential to interact with the vertebrate endocrine system. I encourage the EPA to proceed with putting this assay on-line, while they concurrently address the many concerns raised in Items 2 and 4 above.

Additional Comments and Materials Submitted

David Crews: References

Cooke, P.S. (1996). Thyroid hormone and regulation of testicular development. Anim. Reprod. Sci. 42: 333–341.

Crews, D., Willingham, E., Skipper, J.K. (2000). Endocrine disruptors: Present issues, future directions. Quart. Rev. Biol. 75: 243-260.

Fort, D.J., Degitz, S., Tietge, J., Touart, L.W. (2007). The hypothalamic-pituitary-thyroid (HPT) axis in frogs and its role in frog development and reproduction. Crit. Rev. Toxicol. 37: 117-1161.

Franca, L.R., Hess, R.A., Cooke, P.S., Russell, L.D. (1995). Neonatal hypothyroidism causes delayed Sertoli cell maturation in rats treated with propylthiouracil: Evidence that the Sertoli cell controls testis growth. Anat. Rec. 242: 57–69.

Hayes, T.B. (1995). Interdependence of corticosterone and thyroid hormones in larval growth and development in the western toad (*Bufo boreas*): I. Thyroid hormone dependent and independent effects of corticosterone on growth and development. J. Exp. Zool. 271*:* 95–102.

Hayes, T.B. (1997a) Steroids as potential modulators of thyroid hormone activity in anuran metamorphosis. Am. Zool. 37: 185–194.

Hayes, T.B. (1997b). Hormonal mechanisms as developmental constraints on evolution: Examples from the Anura. Am. Zool. 37*:* 482–490.

Hayes, T.B. (1998). Sex determination and primary sex differentiation in amphibians: genetic and developmental mechanisms. J. Exp. Zool. 281: 373-399.

Hayes, T.B., Licht, P. (1993). Metabolism of exogenous steroids by anuran larvae.
Gen. Comp. Endocrinol. 91: 250-258.

Kirby, J.D., Jetton, E.A., Cooke, P.S., Hess, R.A., Bunick, D., Ackland, J.F., Turek, F.W., Schwartz, N. (1992). Developmental hormonal profiles accompanying the neonatal hypothyroidism-induced increased in adult testicular size and sperm production in rat. Endocrinology 131: 559–565.

Matta, S.L.P, Vilela, D.A.R. Godhinho, H.P., Franca, L.R. (2002) The goitrogen 6-*n*-propyl-2-thiouracil (PTU) given during testis development increases sertoli and germ cell numbers per cyst in fish: the tilapia (*Oreochromis niloticus*) model. Endocrinology 143: 970–978.

Mosconi, G. DiRosa, I., Bucci, S. Morosi, L. Franzoni, M.F., Polzonetti-Magni, A.M., Pascolini, R. (2005). Plasma sex steroid and thyroid hormones profile in male water frogs of the *Rana esculenta* complex from agricultural and pristine areas. Gen Comp Endocrinol. 142: 318-24.

Ng, S.M., Wong, S.C., Isherwood, D.M., Didi, M. (2007). Biochemical severity of thyroid ectopia in congenital hypothyroidism demonstrates sexual dimorphism. Eur. J. Endocrinol. 156:49-53.

Schulz, R.W., Menting, S. Bogerd, J. Franca, L.A. Daniel A.R. Vilela, D.A.R. Godinho. H.P. (2005). Sertoli cell proliferation in the adult testis—evidence from two fish species belonging to different orders. Biol. Reprod. 73: 891–898.

Catherine Propper: Reference List

Buchholz DR, Paul BD, Fu L & Shi YB 2006 Molecular and developmental analyses of thyroid hormone receptor function in Xenopus laevis, the African clawed frog. *Gen.Comp Endocrinol.* **145** 1-19.

Gray KM & Janssens PA 1990 Gonadal hormones inhibit the induction of metamorphosis by thyroid hormones in Xenopus laevis tadpoles in vivo, but not in vitro. *Gen.Comp Endocrinol.* **77** 202-211.

Pfaff DW, Vasudevan N, Kia HK, Zhu YS, Chan J, Garey J, Morgan M & Ogawa S 2000 Estrogens, brain and behavior: studies in fundamental neurobiology and observations related to women's health. *J Steroid Biochem.Mol.Biol.* **74** 365-373.

Vasudevan N, Ogawa S & Pfaff D 2002 Estrogen and thyroid hormone receptor interactions: physiological flexibility by molecular specificity. *Physiol Rev.* **82** 923-944.

Richard Wassersug: REFERENCES

Degitz, S.J., P.A. Kosian, E.A. Makynen, K.M. Jensen and G.T. Ankley 2000 Stage- and species-specific developmental toxicity of all-trans retinoic acid in four native North American ranids and *Xenopus laevis*. *Toxicol. Sci*., 57:264-274.

Degitz, S.J., E.J. Durham, J.E. Tietge, P.A. Kosian, G.W. Holcombe and G.T. Ankley 2003 Developmental toxicity of methoprene and several degradation products in *Xenopus laevis*. *Aquat. Toxicol.*, 64:97-105.

Denver, R.J. 1996 Neuroendocrine control of amphibian metamorphosis. In: "*Metamorphosis: Post-Embryonic Reprogramming of Gene Expression in Amphibian and Insect Cells*." J. R. Tata, L. I. Gilbert and E. Frieden (eds.) Academic Press, Orlando, pp. 433-464.

Dubois, G.M., A. Sebillot, G.G.J.M. Kuiper, C.H.J. Verhoelst, V.M. Darras, T.J. Visser and B.A. Demeneix 2006 Deiodinase activity is present in *Xenopus laevis* during early embryogenesis. *Endocrinology*, 147:4941-4949.

Feder, M.E. and R.J. Wassersug 1984 Aerial versus aquatic oxygen consumption in larvae of the clawed frog, *Xenopus laevis*. *J. Exp. Biol.*, 108:231-245.

Feder, M.E., D.B. Seale, M.E. Boraas, R.J. Wassersug and A.G. Gibbs 1984 Functional conflicts between feeding and gas exchange in suspension-feeding tadpoles, *Xenopus laevis*. *J. Exp. Biol.*, 110:91-98.

Fort, D.J., S. Degitz, J. Tietge, L.W. Touart 2007 The hypothalamic-pituitary-thyroid (HPT) axis in frogs and its role in frog development and reproduction. *Crit. Rev. Toxicol.*, 37:117-161.

Gardiner, D., A. Ndayibagira, F. Grun and B. Blumberg 2003 Deformed frogs and environmental retinoids. *Pure Appl. Chem*., 75:2263-2273.

Hayes, T.B., R. Chan and P. Licht 1993 Interaction of temperature and steroids on larval growth, development and metamorphosis in a toad (*Bufo boreas*). *J. Exp. Zool*. 266:206-215.

Hayes, T.B. 1997 Steroid as potential modulators of thyroid hormone activity in anuran metamorphosis. *Am. Zool*., 37:482-490.

Hoff, K. and R.J. Wassersug 1986 The kinematics of swimming in larvae of the clawed frog, *Xenopus laevis*. *J. Exp. Biol*., 122:1-12.

Katz, L., M. Potel and R.J. Wassersug 1981 Structure and mechanisms of schooling in larvae of the clawed frog, *Xenopus laevis*. *Anim. Behav.*, 29:20-33.

Lum, A., R.J. Wassersug, M. Potel and S. Lerner 1982 Schooling behavior of tadpoles: A potential indicator of ototoxicity. *Pharmac. Biochem. Behav*., 17:363-366.

Maden, M. 1993 The homeotic transformation of tails into limbs in *Rana temporaria* by retinoids. *Dev. Biol.*, 159:379-391.

Nishikawa, K. and R.J. Wassersug 1988 Morphology of the caudal spinal cord in *Rana* (Ranidae) and *Xenopus* (Pipidae) tadpoles. *J. Comp. Neurol*., 269:193-202.

Ortiz-Santaliestra, M.E. and D.W. Sparling 2007 Alteration of larval development and metamorphosis by nitrate and perchlorate in Southern Leopard Frogs (*Rana sphenocephala*). *Arch. Environ. Con. Tox.*, 53:639-646.

Pronych, S. and R.J. Wassersug 1994 Lung use and development in *Xenopus laevis* tadpoles. *Can. J. Zool*., 72:738-743.

Robins A., G. Lippolis, A. Bisazza, G. Vallortigara and L.J. Rogers 1998 Lateralized agonistic responses and hindlimb use in toads. *Anim. Behav*., 56:875-881.

Rot-Nikcevic, I., R.J. Denver and R.J. Wassersug 2005 The influence of visual and tactile stimulation on growth and metamorphosis in anuran larvae. *Funct. Ecol*., 19:1008-1016.

Seale, D.B. and R.J. Wassersug 1979 Suspension feeding dynamics of anuran larvae related to their functional morphology. *Oecologia*, 39:259-272.

Seale, D.B., K. Hoff and R.J. Wassersug 1982 *Xenopus laevis* larvae (Amphibia, Anura) as model suspension feeders. *Hydrobiologia*, 87:161-169.

Söderman, F., S. Dongen, S. Pakkasmaa and J. Merilä 2007 Environmental stress increases skeletal fluctuating asymmetry in the moor frog *Rana arvalis*. *Oecologia*, 151:593-604.

Vershinin, V.L., E.A. Gileva and N.V. Glotov 2007 Fluctuating asymmetry of measurable parameters in *Rana arvalis*: Methodology. *Russ. J. Ecol+.,* 38:72-74.

Wager, V.A. 1965 *The Frogs of South Africa*, Purnell & Sons, Johannesburg.

Walks, D.J. 2007 Persistence of plankton in flowing water. *Can. J. Fish. Aquat. Sci*., 64:1693-1702.

Wassersug, R.J. and M.E. Feder 1983 The effects of oxygen concentration, body size and respiratory behaviors on the stamina of obligate aquatic (*Bufo americanus*) and facultative air-breathing (*Xenopus laevis* and *Rana berlandieri*) anuran larvae. *J. Exp. Biol*., 105:173-190.

Wassersug, R.J. 1996 The biology of *Xenopus* tadpoles. In: R.C. Tinsley and H.R. Kobel (eds.). *The Biology of Xenopus*, Clarendon Press, Oxford, pp. 195-211.

Wassersug, R.J. 1989 Locomotion in amphibian larvae (or "Why aren't tadpoles built like fishes?"). *Amer. Zool*., 29:65-84.

Wassersug, R.J. and A.M. Murphy 1987 Aerial respiration facilitates growth in suspension-feeding anuran larvae (*Xenopus laevis*). *Exp. Biol*., 46:141-147.

Wassersug, R.J. 1997 Where the tadpole meets the world-Observations and speculations on biomechanical and biochemical factors influencing metamorphosis in anurans. *Amer. Zool*., 37:124-136.

Wells, K.D. 2007 *The Ecology and Behavior of Amphibians*, University of Chicago Press, Chicago, pp. 608.

Shi, Y. 2000 *Amphibian Metamorphosis*, John Wiley & Sons, Inc., Toronto.

PEER REVIEW COMMENTS ORGANIZED BY REVIEWER

Peer review comments received for the amphibian metapomphosis assay are presented in the sub-sections below and are organized by reviewer. Peer review comments are presented in full, unedited text as received from each reviewer.

David Crews Review Comments

REPORT ON THE AMPHIBIAN METAMORPHOSIS ASSAY (AMA) AS A POTENTIAL SCREEN IN THE ENDOCRINE DISRUPTOR SCREENING PROGRAM (EDSP) TIER-1 BATTERY

DAVID CREWS
UNIVERSITY OF TEXAS AT AUSTIN

OVERVIEW

There is much to praise about this report, in particular the careful thought and precision of the experimental protocol in all three phases of the process. However, it is the opinion of this reviewer that the conclusions regarding inter-laboratory variability are not warranted and that it fails as a method for accomplishing the stated goal of the assay to be part of the Endocrine Disruptor Screening Program (EDSP). This assessment is based on the fact that endocrine disrupting compounds are rarely (if ever) found in nature as the sole contaminant, that such mixtures interact in a manner that must be tested before the interactions can be discarded as factors, and that endocrine disrupting compounds/chemicals (EDCs) act on integrated endocrine systems during development that have consequences beyond the life history of the individual organism. As a traditional environmental toxicology exercise, the assay is a first step, but still ignores the issue of low dosages and the need for other endocrine endpoints.

REVIEW

1. Clarity of the stated purpose of the assay.

The documents provided document the rationale for an amphibian metamorphosis assay (AMA) as a high throughput *in vivo* assay for thyroid disrupting chemicals. A series of tests designed to validate this method are described using the tadpole *Xenopus laevis*.

The document "Integrated Summary Report – Amphibian Metamorphosis Assay" (File Name: Ama_isr) presents a protocol designed such that an aquatic toxicology laboratory would be able to conduct studies of chemicals for their effects on the developing thyroid system of this animal model system.

Specifically, tadpoles reared under standardized conditions will be treated during a discrete period of development beginning at Stage 51 will be exposed for 21 days to one of several concentrations of the test chemical; another group will be exposed to a water control. Within each chemical treatment there will be four replicates. At each of three time points (d0, d7 and d21 or treatment) the endpoints measured will include developmental stage, wet weight, snout-to-vent length (SVL), whole body length (WBL), hind limb length (HLL), and thyroid histology. The latter two measurements will utilized dissecting (limb length) or light microscopic measurements with computer-assisted image-digitizing software measurements. Finally, tadpoles will be observed daily for mortality and malformations.

A flow-through method for delivering the chemicals at the various concentrations will be used with measurements being taken at periodic intervals (weekly) to evaluate and validate the composition of the water. It appears that the preferred system will require that each set of 4 replicate tanks (= test vessel) will

receive a given concentration using a diluter system. It is commendable that in this method the test tank will not serve as a feed to other tanks. The alternative method, static-renewal, is not described and so cannot be evaluated by this reviewer.

Each replicate tank will be a 4 litre glass aquarium with 20 larvae initially. Light, temperature, pH, DO, and feeding will be standardized, with the tanks randomly situated to allow for possible differences due in placement.

Adult male and female South African clawed frog *Xenopus laevis* will be injected in human chorionic gonadotropin (hCG) to induce breeding. The source of the adult animals (pg. 5), and the "best spawns" (pg. 5) are a concern (see 2. below). The larvae will be raised in constant densities, being fed twice daily during the week and once daily on weekends and holidays.

Three test concentrations will be utilized. The highest, the maximum test concentration (MTC), is defined as the highest test concentration of the chemical that results in less than 10% acute mortality. This is a concern (see below). The lower concentrations to be tested would be calculated as a dose separation of 0.33-0.5 (max-min).

Test animals will be selected on the basis of normal body morphology and using the hind limb morphology staging criteria of Nieuwkoop and Faber (pg. 10). For a d0 measure, approximately 20 individuals will be measured for WBL. It is not clear if these 20 individuals will be reintroduced into the test population for distribution into the tanks, or whether they will be used to obtain the other stated measurements (see above).

A sample of 5 tadpoles will be taken from each tank on d7, for a total sample size of 20 tadpoles for each treatment/dose, and a detailed selection procedure is outlined for obtaining a similarly sized sample on d21.

The histological measurements are described well as are the statistics to be applied. Procedures for Data Reporting are similarly clear, but should be made mandatory, rather than recommended. It is not clear what is meant by "gross deviations from the test method" and so cannot be evaluated. This should be rigorously defined.

The document "Guidance Document on Amphibian Thyroid Histology Part 1: Technical guidance for morphologic sampling and histological Preparation" (File Name: AMA_Test_Method_Appendix_1), is overall outstanding in its instructional clarity regarding the handling and euthanasia of tadpoles, biometry, and preparation of the samples for analysis. As one who has over 35 years of experience in all aspects of histology, especially paraffin processing, I cannot think of anything that has not been anticipated. There is one important factor that is omitted, however, is consideration of asymmetrical limbs (see below).

The document "Guidance Document on Amphibian Thyroid Histology Part 2: Approach to reading studies, diagnostic criteria, severity grading, and atlas" (File Name: AMA_Test_Method_Appendix_2), is also outstanding in its instructional clarity, breadth and depth. I am impressed by the careful attention to avoiding errors in assessment in addition to the more standard diagnostic criteria for grading slides. The section images themselves are outstanding in both magnifications and the histologist who prepared them is to be congratulated. However, there is a serious flaw in **Section IB. Approach to reading studies** (see below) regarding the scoring of the slides.

2. Clarity, comprehensiveness and consistency of the data interpretation with the stated purpose of the assay.

As instructed (**pg. 12**) in the document "FINAL REPORT OF THE VALIDATION OF THE AMPHIBIAN METAMORPHOSIS ASSAY FOR THE DETECTION OF THYROID ACTIVE SUBSTANCES: PHASE 1: OPTIMISATION OF THE TEST PROTOCOL" (File Name: OECD_Phase_1_Report.pdf), this report and that following "FINAL REPORT OF THE VALIDATION OF THE AMPHIBIAN METAMORPHOSIS ASSAY: PHASE 2: MULTI-CHEMICAL INTERLABORATORY STUDY" (File

Name: OECD_Phase_2_Report.pdf) will be considered together. However, I will focus on the data from the stage 51, 21 d treatment group for Phase 1 since that is the developmental stage for the initiation of treatment in the AMA protocol.

PHASE 1 REPORT

Summary *i)* It is stated that the origin of the effort to develop and optimize an AMA "originated at a meeting of the Amphibian Expert Group, an advisory group to the Validation Management Group, in June 2003 at a meeting hosted by the US Environmental Protection Agency in Duluth, MN, USA." (pg. 18) There is no reference to another EPA-sponsored workshop (DeVito et al., 1999). This is unfortunate because a specific observation/caution was made (Thyroid function affects reproductive development and function). Further, a specific recommendation appears to have been ignored in the present effort. Namely, "A number of assays or test systems can be used to detect chemicals that produce hypothyroidism. However, most of these assays or test systems are time consuming and not necessarily specific for hypothyroidism. In addition, pronounced decreases in serum T4 concentrations are required to detect the behavioral or morphologic changes. Alterations in serum THs can be detected at lower dose levels than those required to detect the behavioral and morphologic changes in these systems. Because of the greater sensitivity and simplicity, determination of serum TH concentrations is recommended instead of these developmental assays. It should be remembered that using adult, pubescent, or prepubescent animals may be qualitatively predictive of fetal response, whereas it may not be quantitatively predictive of dose or response in fetal tissue." (pg. 412 of DeVioto et al., 1999).

Summary *iv* and *vi)* In the first phase three participating laboratories each used "their specific methods to test the anti-thyroid compound, 6-propylthiouracil (PTU), and the receptor agonist, T4, at comparable exposure concentrations." In the second phase identical methods were used by six participating laboratories with a total of 14 experimeintal studies with the replication of T4, and two new chemicals, specifically sodium percholorate (Na-PER), a thyroid hormone synthesis inhibitor, and iopanoic acid (NIS), a deiodinase inhibitor.

Statistical Analysis (pg. 23, Phase 1).

Gene expression (item 14). It is not appropriate to simply presume that the gene expression data followed a log-normal distribution. It should first be tested for heterogeneity of variance and then, if appropriate, the transform done. Further, there description of the methodology for the semi-quantitative RT-PCR ("densitometric analysis of scanned agarose gels are shown. Results were expressed relative to the control Group") is not adequate. Show me the protocol and the original data so that I can determine the validity of the method.

Analytic Chemistry Results Standard Deviations (Tables 3 and 6). A replicate is defined on Pg. 22 and described as "20 tadpoles were used per replicate tank in the GER and JPN laboratories; the US laboratory used 25 tadpoles per replicate tank in the PTU studies." Considering only the Stage 51 sutdy, the variability in PTU concentrations in the US laboratory is commendable, but that of the JPN laboratory is of concern. This is amplified in the lack of a 0.00 concentration in the JPN measurements, raising the question of whether their control water actually has compounds that cross react in the measurement system. A similar problem exists for the T4 concentrations (Tables 4 vs. 7).

Item 21. Comparison of Control Data (pg. 25). This is a misrepresentation as there is no data provided by the GER laboratory, and that of the JPN laboratory is questionable.

Table 8. Consideration of the median is misleading in that the tadpoles from the GER laboratory have a bimodal distribution of development for the PTU, and develop slower under the T4 regimen.

Table 11. It is extremely odd that in the JPN laboratory control tadpoles from the two treatment groups varied substantially (there is no overlap by one STD).

Item 27 and Table 12 (pgs. 28 and 29). "The significant difference at 5 mg/L after 14 days of exposure in JPN study seems to be an anomalous result and driven by one of the two replicates which does not fit the pattern of the other tests." Considering the above comment under Item 21, this may not be so anomalous and should not be disregarded. It is this reviewer's opinion that the absence of analytic chemistry of the GER lab, and the questionable quality of the analytic chemistry of the JPN lab, there is really no points of comparison from the null condition.

Table 13 vs. Table 15 Comparison. These tables present data from two laboratories (GER and JPN, respectively) for the same treatment conditions. However, if we compare the information for hind limb length for the GER lab, we see that the difference in Pool means between d7 and d21 values are:

	Control	**2.5 mg/L**	**5.0 mg/L**	**10 mg/L**	**20 mg/L**
GER	8.9	7.6	6.6	6.8	6.4
JPN	10.6	11.4	8.7	10.2	3.2

It does not take a scientist to come to the conclusion that the data produced by the two laboratories are not comparable.

Figure 2. The substantial SEMs in the d21 TSHb and BTEB are troublesome.

Tables 27, 29 and 31 Comparison. The only valid measure of inter-laboratory concordance is that of body weight. Comparing the difference between the control and the 2.0 mg/L T4 average values for the GER, JPN, and US sites are: 274, 205, 402, respectively, this and inspection of the trends within each lab, I conclude that they cannot be compared.

Conclusion: Phase 1 data is not valid in terms of inter-laboratory comparison. While **"**these studies resulted in remarkably similar outcomes among the different laboratories, despite minor methodological differences", the results from each laboratory cannot be combined one with the other, severely limiting any attempts at meta-analysis.

PHASE 2 REPORT

Summary (pg. 19). The purpose of the " Phase 2 of the validation study aimed at an inter-laboratory multi-chemical testing with an harmonised protocol." Specifically, tadpoles reared under standardized conditions were treated during a discrete period of development beginning at Stage 51 for 21 days to one of concentrations of the test chemical; another group will be exposed to a water control. Within each chemical treatment there will be four replicates. At each of three time points (d0, d7 and d21 or treatment) the endpoints measured will include developmental stage, body mass as wet weight, SVL, WBL, and HLL as well thyroid histology. Six international laboratories performed a total of 14 studies using Na-PER (n=4) and IOP (n=4).

Identity of Laboratories. This information is not provided and this is very regrettable. It is vital to know if any given laboratory can reproduce its data for certain controls, in this instance the no chemical group and the T4 group. If one or more of the three laboratories in the Phase 1 study participated in the Phase 2, this would enable evaluation of QC/QA. This point is further evidenced in the finding (Tables 8 and 9) that "The intra-laboratory comparison of tadpole growth parameters showed highly reproducible results in lab 1, lab 2 and lab 5 and less reproducible results in lab 3." (pg. 34)

Growth in the Control Group (pp. 33-40) While reassuring, the finding that tadpole growth within the

control groups within a particular laboratory are reproducible, this is not at all satisfactory if the aim is to be able to compare across laboratories. Table 9 in particular would convince any reviewer for a reputable scientific journal to recommend rejection.

Effects of Na-PER and IOP on Developmental Endpoints. If it is not possible to compare the laboratories in terms of the control group, then there is no point in attempting to make sense of the inter-laboratory variation in the experimental groups. In this regard, lab 1 has a reasonable dose-response curve for Na-PER at d21 for WBL, SVL, and mass (PER (Tables 12-14).

Items 55, 64 and 77. The presentation of results for histopathology of two laboratories that are not comparable is misleading at best. What kind of conclusion can be drawn from this data?

Effects of T4 on Developmental Endpoints. The comparison of Tables 23-26 suggest that tadpoles in laboratories 1 and 2 showed limb growth but little or no change in mass, whereas animals in , the animals limbs responded but not mass, whereas for laboratories 3 and 4 the opposite pattern existed.

Conclusion. Phase 2 was conducted in an exemplary fashion in terms of standardizing protocols. The conclusion that "these studies resulted in remarkably similar outcomes among the different laboratories, despite minor methodological differences" is certainly true within each laboratory. However, the results from each laboratory cannot be combined one with the other, severely limiting any attempts at meta-analysis. Thus, the most important opportunity this Phase allowed, namely the comparison across laboratories, is an unqualified negative. Finally, it is vital that any laboratories that participated in both Phases 1 and 2 be compared for control group measurements.

PHASE 3 REPORT

Summary. In the Phase 3 study additional compounds were recommended for study, benzophenone-2 (BP-2), 17β-estradiol (E2), potassium iodide (KI) and p,p'-DDE (DDE), but experiments were only conducted on BP-2 and E2. However, the concept of including both positive and negative controls in Phase 3 is excellent.

Control group. Inspection of Table 2 and the statement on pg. 16 "there was no solvent control" suggesting there was no control group in this study. If this indeed were the case, then no conclusions can be drawn about the relationship between BP-2 and E2. This clearly is an omission, but an important one in Table 2..

Statistical Analysis. If there is no control group, what is the basis of the statement on pg 16 "Dunn's test was used for pairwise comparisons of treatment group medians to the control median" and on pg. 17 "pairwise comparisons of treatment group means to the control mean were performed using Dunnett's/Tamhane-Dunnett's test."

Growth in the Control Group. As stated (pg. 22) "Control tadpoles used in the two independent experiments in lab 1 showed similar growth rates indicating low intra-laboratory variability. In comparison to lab 1, control tadpoles used in the experiment performed in lab 2 were greater in size, as judged from WBL and SVL measurements on day 7, and had increased body weights." However, Tables 3-5 do not contain data clearly labeled as the 0.0 or DWC group. While this can be understood for Table 3 (as it is d0), the legends for Tables 4 and 5 as well as for Figures 2 and 3 caused this reviewer considerable time and effort before understanding that they were misstated.

Sex Determination. It is not clear how sex assignment was determined. What were the criteria used in the "gross morphological assessment" (pg. 26)?

Table 7 (pg. 28). It should be noted that the effect of the 2.0 and 10 mg/L E2 is most likely due to the reduction in the variance, which almost certainly is due to the larger sample size.

Discussion. The interaction of the thyroid system is presented as unidirectional and cause-effect (pg. 51), that is how gonadal steroid hormones affect the pituitary-thyroid axis or with TH action. This is misleading. First, it does not consider how the thyroid might affect the developing gonad. Second, the emphasis should be on the interaction of two axes during development, namely the hypothalamo-pituitary-gonad and the hypothalamo-pituitary-thyroid axes.

It is understood that interpretations of the literature are prone to the biases of the reviewer. This reviewer disagrees with the statement of the authors of Phase 3 that "interference of gonadal steroids with the thyroid system occurs most likely at the hypothalamic-pituitary level" (pg. 52) and present additional evidence in the section 5. Limitations. f.

Conclusion. Overall a good study. Consideration however should be given to the issues identified above.

3. Biological and toxicological relevance of the assay as related to its stated purpose.

This assay is designed as a standard toxicological screen. As such, it accomplishes its goal. However, a number of studies have now shown unequivocally that traditional toxicological studies are ill-suited for detecting chemicals that have endocrine disrupting capacity. There are multiple reasons for this and are listed below.

a. The thyroid system is part of an integrated endocrine system that is essential not only for normal functioning at particular life stages, but also for advancing the developing organism through a series of carefully regulated stages that result in a functional (= reproductive adult). Hence, it cannot be considered in isolation of other endocrine systems. This is particularly the case when considering the developing reproductive system.

The present document describes the effects of compounds on the thyroid axis and its consequences on limb growth. This ignores the fact that factors influencing the thyroid axis may also affect the reproductive axis. For example, a recent study comparing populations in frogs in a contaminated (by agricultural runoff) and a pristine lake in Italy document that the pattern of circulating concentrations of steroid hormones and T3 and T4 are disrupted and the testes of adults affected (Mosconi et al., 2005). In laboratory experiments administration of goiterogens such as thiourea and 6-*n*-propyl-2-thiouracil (PTU) can alter normal patterns of sex determination in Xenopus and other frogs as well as fishes and mammals (Fort et al., 2007; Franca et al., 1995; Hayes, 1997a, b; Matta et al., 2002; Schultz et al., 2005). Significantly, after treatment is stopped, spermatogenesis is restored to normal levels (Cooke, 1996; Kirby et al., 1992; Schultz et al., 2005). Thus, such compounds lead to increased interstitial cell growth and activity, to the extent that in the male spermatogenesis is inhibited (presumably due to an overproduction of androgen). Such studies indicate that thyroid hormone is important in normal gonadal development and, further, that interference at this level will produce sterile individuals.

b. EDCs are ubiquitous in natural environments. Standard toxicological screening methods have a focus of determining whether a given compound is toxic, leading to death (defined here by the LC 50 and/or to body and organ malformations). Two typical life stages in which compounds are tested are the adult or developing (embryonic or early life) organism. In both instances the emphasis is on the individual organism within a single generation. In addition, any number of compounds when administered to developing organisms may have no demonstrable effect on mortality or growth. However, these compounds, and particularly EDCs, can affect sexual development—even at extremely low doses. Such sterile individuals occupy space and use resources but cannot contribute to the growth of the population, as their genes will not transmit to subsequent generations, hence leading to an evolutionary death.

4. Clarity and conciseness of the test method in describing the methodology of the assay such that a laboratory can:

a. comprehend the objective: objectives stated in AMA Test Method and Appendices (**File names:** AMA_Test_Method, Appendix_1; and Appendix_2) were clear and concise. Table 3 should also include daily observation of gross morphological deformities to be consistent with text (File name**:** AMA_Test_Method, pg. 8)

b. conduct the assay: Methods and materials in the documents mentioned above were detailed.

c. observe and measure prescribed endpoints: Pictoral references for histology readings, morphological measurements and image set up in AMA Test Method Appendices allow adequate standardization of measurements among multiple laboratories.

d. compile and prepare data for statistical analyses: Please see previous comments on statistical analysis for PHASE 1 and PHASE 3 in section 2 of this review.

e. report results: Performance criteria described in Table 4 of AMA Test Method (File name: AMA_Test_Method, pg. 14) provided detailed requirements of reportable data. Concern of small sample size at d21 compounded with mortality at this stage, as well as varying developmental stages within one treatment tank should be addressed. This reviewer did not evaluate the alternate static renewal design.

If warranted, please also make suggestions or recommendations for test method improvement.

a. Static-renewal. The alternative method, static-renewal, is described for insoluble compounds and high concentrations relative to the limits of water solubility (File name: Battelle_multi-chem_report) but is not used in subsequent Phase studies as the chemicals tested were water soluble. However, it is not described and so cannot be evaluated. If static renewal refers to the regular (periodic) replacement of the water in the tank, this is fraught with difficulties, not the least of which is the buildup of metabolic byproducts that can affect the endocrinology of the animals. Finally, if the Phase 1-3 testing was conducted using flow-through systems, alternatives such as static renewal should be disallowed until comparable tests for intra- and inter-laboratory QA/QC are conducted.

b. Source of animals and selection of spawns. The source of the adult animals (pg. 5), and the "best spawns" (**pg. 5**) are a concern. That is, it appears that all of the egg masses will be collected together and a selection is made. This could result in only a few of the mating pairs producing most of the tadpoles used in any specific study. This may be mitigated by the treatment of the selected spawns being treated with a 2% L-cystein solution and then combining the larvae, but it would be preferable to use ALL spawns produced treat them all, and selecting the larvae after they are freed from the jelly coat. The large discrepancy in the animals in the control groups in Phase 2 illustrates the importance of the source of animals.

c. Analysis of food. The quality control (QC) of the food offered to the larvae/tadpoles are not described and are of concern. Is there documentation and analytic verification available for each production? While the same vendor is being used (Sera GmbH), it is well known that batches of commercially available foods for laboratory animals can vary significantly. Further, if the food is produced by multiple facilities of the same company, and thus purchased by different testing facilities, this can be a significant source of variation between testing facilities.

d. Maximum Test Concentration. The highest test concentration, or MTC, is defined as the highest test concentration of the chemical that results in less than 10% acute mortality (pg. 7). This is a concern. It is stated that if prior empirical acute mortality data are not available or sufficient information is not available to develop regression models to estimate the MTC, then a 96 hr LC50 test will be conducted. The LC50 traditionally is defined as the lowest concentration that results in 50% mortality, but it is not clear if this is

how it is defined here. If, however, this is the definition used here, then the MTC would be calculated as being 1/3 of the LC50. The lower concentrations to be tested would be calculated as a dose separation of 0.33-0.5 (max-min). This does not correspond to best practice NOAEL calculations.

e. Dilutions: Given the concern about low dosage effects, it is not clear why the AMA advises that only three dosages of the test chemical be used. This is particularly puzzling when in the Phase studies four or more dosages, spanning a full log unit or more, were used.

f. Initial sample. For a d0 measure, approximately 20 individuals will be measured for WBL. It is not clear if these 20 individuals will be reintroduced into the test population for distribution into the tanks, or whether they will be used to obtain the other stated measurements (see above).

g. Sample size. A sample of 5 tadpoles will be taken from each tank on d7, for a total sample size of 20 tadpoles for each treatment/dose. This allows for up 15 remaining tadpoles per tank for a second terminal sample on d21 (or a total of 60 tadpoles for each treatment/dose if there is no death or disability). This is unlikely to be the case, and the issue of how the requisite 20 individuals will be selected for in-depth analysis for the d21 sample is considered. If size matched samples are to be used as stipulated, why was the most advanced stage selected for analysis? Also, why is this same criterion not applied to the d7 sample as a distribution of stages are likely to be present as well (although perhaps not as wide a range)?

h. Asymmetrical limbs. The body plan of most animals, including frogs, is not symmetrical. Although differences can be slight, they are present and have been shown in various studies to be important mechanistically as well as evolutionarily. One side should be selected and it be mandated that this side only be measured.

i. Scoring of slides. The Phase I and Phase 2 studies addressed the issue of intra- and inter-laboratory variability. Although the results of both sets of studies indicated this variation to be minimal, with the "response profiles of the various endpoints were different for the individual test substances but reproducible across laboratories", this reviewer still has a concern that any initial screen be conducted in a non-blinded fashion, this must be limited to evaluation of the quality of the sections and their suitability for measurement. It is mandatory that "any potential compound-related findings will be re-evaluated by the pathologist in a blinded manner prior to reporting such findings" (pg. 6). The following terminology "when appropriate" is absolutely inappropriate. The caveat that "Certain diagnostic criteria, such as thyroid gland hypertrophy or atrophy, cannot be read in a blinded manner due to the diagnostic dependence on control thyroid glands" (pg. 6) can be mitigated by having a set of standard slides that are distributed to all potential contractors. The images provided in this document are excellent and could serve this purpose at least initially. Finally, there is a need to assess inter-observer reliability both within the same laboratory as well as across contract laboratories. There should be separate Quality Assurance/Quality Control (QA/QC) performance guidelines.

j. Radioimmunoassay. As stated by DeVito et al. (1999) above, a less costly and time-consuming alternative is available. In these instances, whole bodies or heads can be extracted and TH concentrations can be assayed using either radiometric or ELISA methods.

5. Strengths and/or limitations of the assay.

What follows refers only to the primary document, Test Method for the Amphibian Metamorphosis Assay **(**File names**:** AMA_Test_Method; AMA_Test_Method_Appendix_1; AMA_Test_Method_Appendix_2)

Strengths of the assay

a. It is commendable a flow-through method is recommended. This avoids the problem of buildup of metabolic byproducts that can influence the stated endpoints as may occur in the static-renewal method.

b. The use of widely accepted developmental staging for *Xenopus* development.

c. The use of a defined time window for exposure.

d. The use of computer-assisted software for microscopic determinations.

e. The use of standardized histological protocol.

f. The use of standard histological slides to facilitate evaluation of thyroid histology.

g. The issue of sample selection for the terminal sample (d21) is considered and detailed.

h. The issue of variation within and across laboratories has been addressed in rigorous manner.

i. The statistical evaluation and power analysis as guiding principles for implementation of the AMA is excellent (File name: Power_Analysis).

Limitations of the assay.

a. Low Dose. Recommend dosages spanning at least one full log unit and having at least four concentrations to determine true nature of the dose-response.

b. Threshold. This protocol does not allow for this important determination.

c. Mixtures. This protocol does not allow for this important determination. "Recent findings of a rather strong activity of BP-2 in *in vitro* assays and the marked difference in the severity of BP-2 effects on the thyroid system in two different laboratories could be interpreted that the actual potency of BP-2 to disrupt thyroid system function is strongly dependent on iodide availability." (pg. 70) (File name: OECD_Phase_3_Draft_Report) It is possible that the potency of other chemicals may depend on differential iodide concentrations.

d. Mortality vs. Evolutionary Death. The present EDSP focuses on the individual in its own lifetime. This is valuable information, but says little about the impact of the chemical on the population through time (proximate or ultimate). One measure is whether an individual will breed. If the individual does breed, but its young do not develop properly and do not breed, than the overall result in terms of the population is the same. If the goal is to have a means of evaluating the impact on compounds that have an impact on thyroid function for wildlife and human health, then it is the latter issue that is pertinent.

e. Sex Differences in Sensitivity. If one goes to PubMed and inputs "sex differences in thyroid function", 132 citations come up in the primary literature. If this is further refined to "sex differences in thyroid function, development" 15 papers are cited. Typical is that of Ng et al., (2007) findings that female infants with thyroid ectopia have significantly higher thyroid stimulating hormone (TSH) concentrations than do males and significantly lower circulating concentrations of plasma T4 were significantly lower than in males. Since the animal is being sectioned for histology, it would be a simple (but adding to expense)

addition to look at the gonads. Given that the tested compounds may also influence differentiation of the gonads (see below), it would also be necessary to use standard genetic markers for sexing the tadpoles (see.

f. Multiple Target Organs. Hyperthyroidism induced by PTU or methimazole, also acts on developing gonad, specifically in males on the Sertoli cells. It has been known since 1925 (Rickey) that thyroidectomy eliminates sexual activity in male rats and in more modern experiments in both mammals and fish the Sertoli cells early in testicular differentiation have abundant receptors that decline markedly after sexual maturation (Cooke, 1996; Kirby et al., 1992; Matta et al., 2002; Schultz et al., 2005). Thus, the observation in Phase 3 that E2 caused male-to-female sex reversal without affecting other measures including the histopathology of the thyroid should be considered seriously.

g. Procedure for Training of Pathologists. Need to assess inter-observer reliability both within the same laboratory as well as across contract laboratories. Best course would be to require that a standard set of slides/images be provided to each contract laboratory and Quality Assurance/Quality Control guidelines be developed and adhered to with no exceptions.

h. Measuring the Same Side. The animal body is asymmetrical and so it would be necessary that the same limb be measured on each tadpole.

i. Sample Sizes. The issue of the selection of tadpoles for the sample dates (d0, d7, and d21) are considered above. Here though I raise another issue. If size matched samples are to be used as stipulated, why was the most advanced stage selected for analysis? Also, why is this same criterion not applied to the d7 sample as a distribution of stages are likely to be present as well (although perhaps not as wide a range)? This is unlikely to be the case, but raises the issue of how the requisite 20 individuals will be selected for in-depth analysis for the d21 sample. Further, what is to happen if mortality and disabilities may be such that adequate animal numbers will be available to obtain a meaningful sample?

j. Standardization of Food vs. Potential EDC Content in Food. Specify parameters of the food by analytic chemical analysis and make each contract laboratory supply documentation of having met these criteria with each report.

6. Impacts of the choice of test substances and methods chosen to demonstrate the performance of the assay.
The choice of test substances and methods were reasonable.

7. Repeatability and reproducibility of the results obtained with the assay, considering the variability inherent in the biological and chemical test methods.
This is a major flaw of the material provided and is detailed in the above comments.

8. Please comment on the overall utility of the assay as a screening tool, to be used by the EPA, to identify chemicals that have the potential to interact with the endocrine system sufficiently to warrant further testing.
Before the AMA can be used as a screening tool that is open to contract laboratories, the issues raised above should be addressed. The bottom line is that the AMA is not suitable as a screening tool for endocrine disrupting compounds.

References

Cooke, P.S. (1996). Thyroid hormone and regulation of testicular development. Anim. Reprod. Sci. 42: 333–341.

Crews, D., Willingham, E., Skipper, J.K. (2000). Endocrine disruptors: Present issues, future directions. Quart. Rev. Biol. 75: 243-260.

Fort, D.J., Degitz, S., Tietge, J., Touart, L.W. (2007). The hypothalamic-pituitary-thyroid (HPT) axis in frogs and its role in frog development and reproduction. Crit. Rev. Toxicol. 37: 117-1161.

Franca, L.R., Hess, R.A., Cooke, P.S., Russell, L.D. (1995). Neonatal hypothyroidism causes delayed Sertoli cell maturation in rats treated with propylthiouracil: Evidence that the Sertoli cell controls testis growth. Anat. Rec. 242: 57–69.

Hayes, T.B. (1995). Interdependence of corticosterone and thyroid hormones in larval growth and development in the western toad (*Bufo boreas*): I. Thyroid hormone dependent and independent effects of corticosterone on growth and development. J. Exp. Zool. 271*:* 95–102.

Hayes, T.B. (1997a) Steroids as potential modulators of thyroid hormone activity in anuran metamorphosis. Am. Zool. 37: 185–194.

Hayes, T.B. (1997b). Hormonal mechanisms as developmental constraints on evolution: Examples from the Anura. Am. Zool. 37*:* 482–490.

Hayes, T.B. (1998). Sex determination and primary sex differentiation in amphibians: genetic and developmental mechanisms. J. Exp. Zool. 281: 373-399.

Hayes, T.B., Licht, P. (1993). Metabolism of exogenous steroids by anuran larvae.
Gen. Comp. Endocrinol. 91: 250-258.

Kirby, J.D., Jetton, E.A., Cooke, P.S., Hess, R.A., Bunick, D., Ackland, J.F., Turek, F.W., Schwartz, N. (1992). Developmental hormonal profiles accompanying the neonatal hypothyroidism-induced increased in adult testicular size and sperm production in rat. Endocrinology 131: 559–565.

Matta, S.L.P, Vilela, D.A.R. Godhinho, H.P., Franca, L.R. (2002) The goitrogen 6-*n*-propyl-2-thiouracil (PTU) given during testis development increases sertoli and germ cell numbers per cyst in fish: the tilapia (*Oreochromis niloticus*) model. Endocrinology 143: 970–978.

Mosconi, G. DiRosa, I., Bucci, S. Morosi, L. Franzoni, M.F., Polzonetti-Magni, A.M., Pascolini, R. (2005). Plasma sex steroid and thyroid hormones profile in male water frogs of the *Rana esculenta* complex from agricultural and pristine areas. Gen Comp Endocrinol. 142: 318-24.

Ng, S.M., Wong, S.C., Isherwood, D.M., Didi, M. (2007). Biochemical severity of thyroid ectopia in congenital hypothyroidism demonstrates sexual dimorphism. Eur. J. Endocrinol. 156:49-53.

Schulz, R.W., Menting, S. Bogerd, J. Franca, L.A. Daniel A.R. Vilela, D.A.R. Godinho. H.P. (2005). Sertoli cell proliferation in the adult testis—evidence from two fish species belonging to different orders. Biol. Reprod. 73: 891–898.

David Furlow Review Comments

1. Clarity of the stated purpose of the assay.

The purpose of the assay is to screen for environmental compounds that affect the hypothalamus-pituitary-thyroid axis, using an intact animal model. Overall the document is clear, and the amount of work setting up and evaluating the system is very impressive. Indeed, a standardized method for raising *Xenopus laevis* through metamorphosis for this level of analysis has been surprisingly lacking. The advantages of the system are clear: the system has dramatic, easily measured external morphological changes to a hormone that is identical in structure to its human counterpart. Furthermore, the assay is conducted in a developing animal as opposed to the other battery of whole animal assays the EPA is considering that are conducted in pubertal or adult rats (pubertal male and female rat assay; ovariectomized female rat assay).

The one statement I would add to the stated purpose section is that the assay can also detect disruption of thyroid hormone signaling at the target cell i.e. the presence of thyroid hormone receptor agonists or antagonists (especially since the recommended starting stage is 51 prior to the precence of detectable circulating TH). As stated, the implication is that the assay will only detect disruption of the pathways controlling thyroid hormone synthesis.

2. Clarity, comprehensiveness and consistency of the data interpretation with the stated purpose of the assay.

The endpoints of the assay are stated as the following: mortality, hindlimb length, whole body and snout vent length, developmental stage (although this is primarily based on fore- and hind-limb size and morphology during the stages analyzed), body weight, and thyroid gland histology.

However, I am concerned that the stage 51, 21 day assay is not sufficiently comprehensive or sensitive to detect interference with the HPT axis. Control animals (both in the Phase I and Phase II trials) only usually progress to stage 58 or 59. This precludes any consideration of compounds that affect tail resorption that demands attainment of the highest levels of T3 in the tissue to respond. As an example, overexpression of prolactin does not inhibit any observable aspect of progression through metamorphosis except for resorption of the connective tissue of the tail (Huang H, Brown DD. Prolactin is not a juvenile hormone in Xenopus laevis metamorphosis. Proc Natl Acad Sci U S A. 2000 97(1):195-9.). Perhaps even more relevant, transgenic overexpression of the Type III deiodinase that degrades T4 and T3 arrests animals between stages 60 and 61 with the most obvious effect on gill and tail resorption (Huang H, Marsh-Armstrong N, Brown DD.Metamorphosis is inhibited in transgenic Xenopus laevis tadpoles that overexpress type III deiodinase. Proc Natl Acad Sci U S A. 1999 96(3):962-7.). Limb growth was not affected in this experiment.

The conclusion that the assay is sufficiently reproducible between laboratories will be addressed under item 7.

3. Biological and toxicological relevance of the assay as related to its stated purpose.

The biological and toxicological relevance is clear: metamorphosis is a strictly thyroid hormone driven event, therefore it is reasonable to assume that alterations in the progression of spontaneous metamorphosis by toxicants are the result of disruption of thyroid hormone synthesis and/or action.

4. Clarity and conciseness of the test method in describing the methodology of the assay such that a laboratory can:

a. comprehend the objective,
b. conduct the assay,
c. observe and measure prescribed endpoints,
d. compile and prepare data for statistical analyses, and
e. report results**.**

Comments on assay method narrative:

c. A major source of uncertainty in my mind lies in the use of flow through versus static renewal systems (p. 2). The static renewal system (for tadpoles) is likely the most popular system in most laboratories working with *Xenopus laevis* tadpoles due to convenience and cost (and it is understood that this system may be the only option in some toxicology studies for chemicals with certain properties). It should be noted that silt-filled, murky ponds are apparently the natural habitat of *Xenopus laevis* rather than fast flowing streams. Furthermore, the flow through system would not permit accumulation of degradates of the test chemical that may occur due to light, hydrolysis, and the animal's own metabolic capacity. Nevertheless, the flow system is understandably preferable due to the higher degree of reproducibility and better control of water quality. According to the ISR, (p. 20) a particular exposure system is not *required* but that the flow through system is *preferred*. It would be important to take perchlorate, for example, and test the flow through versus static renewal systems in the same laboratory using the same spawn. In summary, without such a direct comparison, either a static renewal system should always be used so all chemicals can be tested, or the flow through system should always be used and just exclude chemicals that are not suitable for the assay conditions.

b. The statement about "suitable plastics" for system components that do not compromise the study is not clear: certain plastics can likely be ruled out right away such as those that leach BPA and other known endocrine disrupting chemicals (p. 2).

c. In addition, in the flow through system as described on p. 2, it is important to specify that each of the four replicated doses receive an independent water supply (rather than from the same source split into four in order to serve as four independent samples). This point was not clear in the assay description. Perhaps a diagram can be included to clarify the system design.

d. The protocol should recommend whether to dejelly the eggs of spawns used for the assay rather than leaving that up to the individual investigator (p. 5). Dejellying allows much easier sorting of poorly developing embryos that may compromise the rest of the batch. Thus a recommendation one way or the other should be made.

e. The assumption is that the chow from Sera Micron is consistent from lot to lot, but how is this assessed? Are there any guidelines on expiration date or storage conditions? (p. 6).

e. For vehicle controls, a range of concentrations of the most common (ethanol, DMSO) can be tested in the system for effects on metamorphosis (or lack thereof) to make recommendations to testing laboratories. (p. 7).

f. The choice of dosing regimen is unclear (p. 8). While the determination of the MTC is basically clear (although the description of other means to estimate the MTC is rather convoluted), I see a problem with allowing only three doses to be tested with a dose separation of 0.33 to 0.5. For example, the example given does not even satisfy the requirements of the assay as stated: 0.11 of the highest nominal concentration of 1.0 is only 1/9 of the maximum dose. The risk here is that that the assay may not be able to discriminate between general toxicity and a more specific effect on the thyroid hormone driven metamorphosis that may be revealed at lower doses.

If warranted, please also make suggestions or recommendations for test method improvement.

b. Quantitative PCR for gene expression markers of thyroid hormone action (such as well characterized, broadly expressed TH response genes like TR and TH/bZIP, or markers of disruption of the HPT axis like TSH or NIS) would provide a highly quantitative assay that allows the investigator to assess proper thyroid hormone signaling in specific tissues. This aspect of the metamorphosis system is arguably as well, if not better, developed than for estrogen or androgen action in rodents. Futhermore, newer transgenic models are being developed that provide fluorescent or bioluminescent markers of thyroid hormone action in Xenopus.

(For example: the system being developed by Barbara Demeneix's group and the start-up Watchfrog in France: Turque N, Palmier K, Le Mével S, Alliot C, Demeneix BA. A rapid, physiologic protocol for testing transcriptional effects of thyroid-disrupting agents in premetamorphic Xenopus tadpoles. Environ Health Perspect. 2005 113(11):1588-93; Fini JB, Le Mevel S, Turque N, Palmier K, Zalko D, Cravedi JP, Demeneix BA. An in vivo multiwell-based fluorescent screen for monitoring vertebrate thyroid hormone disruption. Environ Sci Technol. 2007 41(16):5908-14.).

d. Since tail resorption is not an endpoint of the whole animal based assay, tail organ cultures are well established, highly reproducible and quantitative, and dose responsive, and would serve to detect interference of compounds directly at a target tissue.

(For example: Schriks M, Zvinavashe E, Furlow JD, Murk AJ. Disruption of thyroid hormone-mediated Xenopus laevis tadpole tail tip regression by hexabromocyclododecane (HBCD) and 2,2',3,3',4,4',5,5',6-nona brominated diphenyl ether (BDE206) Chemosphere. 2006 65(10):1904-8; Ji L, Domanski D, Skirrow RC, Helbing CC. Genistein prevents thyroid hormone-dependent tail regression of Rana catesbeiana tadpoles by targetting protein kinase C and thyroid hormone receptor alpha. Dev Dyn. 2007 236(3):777-90; Furlow JD, Yang HY, Hsu M, Lim W, Ermio DJ, Chiellini G, Scanlan TS. Induction of larval tissue resorption in Xenopus laevis tadpoles by the thyroid hormone receptor agonist GC-1. J Biol Chem. 2004 279(25):26555-62; Lim W, Nguyen NH, Yang HY, Scanlan TS, Furlow JD. A thyroid hormone antagonist that inhibits thyroid hormone action in vivo. J Biol Chem. 2002 Sep 20;277(38):35664-70.)

5. Strengths and/or limitations of the assay.

Strengths:

a. The assay uses an intact animal model that is highly sensitive to thyroid hormone rather than relying solely on cell lines or biochemical assays to predict effects on animal physiology.
b. Chemical analyses are required to make sure compounds meet nominal values.

c. Careful analysis and maintenance of water quality conditions are described to eliminate non-specific effects on metamorphosis.
d. At least one well documented for histopathological assessment is included for comparison to external morphological changes.
e. Two important issues poorly covered by the Agency to date in toxicity evaluations are both addressed here: thyroid hormone synthesis/action and amphibian biology.

Limitations:

a. There is no (or limited) mechanistic component to the assay. It would not be difficult to incorporated gene expression analyses and hormone measurements to the assay.
In Table 1-1 of the ISR, thyroid hormone receptor binding assays and transcriptional activation assays are not listed as additional tests whereas androgen and estrogen receptor based assays are listed as planned.

b. The animals only develop up to a stage just prior to climax; therefore only acceleration or inhibition of hindlimb growth make up the bulk of the analytical component of the assay not effects on tail resorption which may be more sensitive to perturbations in TH levels.

c. Mixtures effects are not accounted for at all. This issue is something that the EPA should start addressing sooner, rather than later. Is the assay robust enough to detect a reversal of T4 effects by IOP, for example?

d. The effects of selective hormone receptor modulators (eg tamoxifen) can be tissue and even species selective in their actions, and endocrine disrupting chemicalss may well follow suit. In this assay, essentially all of the analysis is focused on the hindlimbs (due to the nature of the developmental staging criteria and the direct hindlimb measurements) and thyroid gland histology. Selective effects in specific tissues could be readily determined by incorporating gene expression analysis.

e. Finally, the suitability of *Xenopus laevis* as a surrogate for other amphibians may be questioned. *Xenopus laevis* is a primitive amphibian that does not have a fully terrestrial adult stage, and is not native to North America. In addition, many studies have differing strains of rats can show wide differences in responses to endocrine disrupting compounds and there is essentially no data to my knowledge about this issue in amphibians.

6. Impacts of the choice of test substances and methods chosen to demonstrate the performance of the assay.

The choice of test compounds is highly appropriate, aside from the previously mentioned limitation on the ability of the assay to detect mixture effects. In the interlaboratory exercise in particular, the choice of perchlorate, T4, and iopanoic acid covers three distinct mechanisms of action is highly appropriate.

7. Repeatability and reproducibility of the results obtained with the assay, considering the variability inherent in the biological and chemical test methods.

One of the major concerns about the assay is the degree of inter-laboratory consistency. The first concern, regarding the variability in the progression through metamorphosis by the controls, appears adequately addressed by the lower amount of feed provided in Laboratory 3 (seen in Tables 5 and 6, Interlaboratory report) . Aside from the general delay in metamorphosis and high degree of variance

seen in animals from that laboratory, the degree of consistency within a given laboratory is in fact quite good and the investigators are to be commended.

The second concern is that while overall trends are observed (ie T4 accelerates, perchlorate and IOP delay), there is surprising inconsistency among the laboratories. For example, only labs 3 and 4 detected a significant effect on hindlimb growth and developmental stage at the two highest levels of perchorate tested by day 21 whereas three other labs did not (Table 15, p. 46, Interlaboratory report). Since laboratory 4 apparently had adequate control animal development this cannot simply be due to feeding differences. While the T4 experiments were more in agreement, in the IOP studies, Laboratory 5 shows no effect at all of IOP at either day 7 or day 21 and Laboratory 3 shows a significant effect at 7 days with regard to hindlimb growth (Table 38 p. 70; Interlaboratory report). Furthermore, it was highly surprising that despite effects on hindlimb growth reported in all laboratories except laboratory 5, no significant effects on NF staging were reported. (Table 37 p69 Interlaboratory report). Also, the progression of control animals through metamorphosis by 21 days was remarkably different in this study (Lab 1 ~58, Lab 2 ~59, Lab 3 60-65, Lab 4 ~59, Lab 5 60-62).

Finally, the summary of the thyroid histopathology results are somewhat confusing. In the ISR, p. 60, Figure 5-1, 100% of all glands from all animals where scored as having follicular cell hyperplasia in laboratory 3 whereas the other laboratories scored a generally increasing dose responsive effect. Indeed, across treatment groups, there is a trend of high incidence of abnormality by laboratory 3. Does this reflect lack of experience of the pathologist with scoring amphibian thyroid glands or in the growth and treatment regimen? It might be useful to have the slides from laboratory 3 scored by pathologists from other laboratories, or to have used one pathologist. This concern is amplified in the T4 responses where there is even greater inconsistency between laboratories.

Based on these observations, the consistency of findings across laboratories remains a major concern for the future viability of the assay system.

8. Please comment on the overall utility of the assay as a screening tool, to be used by the EPA, to identify chemicals that have the potential to interact with the endocrine system sufficiently to warrant further testing.

The assay as designed should be able to detect the presence of that, by themselves, can disrupt the normal progression of metamorphosis, and thus by inference, disruption of some point along the hypothalamic-pituitary-thyroid axis or thyroid hormone activity in peripheral tissues. It is an outstanding first step in developing a whole animal bioassay for thyroid hormone system disruption.

However, while there are many excellent aspects of the study design and presentation, several issues summarized above currently preclude the assay's use as a routine screening assay, most notably the high degree of interlaboratory variability, the lack of assessment of endpoints other than basically hindlimb development and thyroid gland appearance, and the recommended dosing regimen is too narrow to discriminate between general toxicity and specific endocrine disruption.

Catherine Propper Review Comments

General Overview:

This assay was developed to determine whether compounds to be testing for Tier 1 Level analysis in the EPA's Endocrine Disruptor Screening Program disrupt thyroid hormone function. Amphibians are an outstanding model for investigations of thyroid hormone function because the process of metamorphosis is strongly regulated by first the expression of the thyroid hormone receptor and then later the secretion of thyroid hormones from the thyroid gland. Therefore, compounds that mimic thyroid hormone activity may increase the rate of metamorphosis, and those that antagonize thyroid hormone activity or function can decrease the rate of metamorphosis. Clear morphological and developmental endpoints are readily evaluated to determine the impact of exposure. Therefore this assay is readily transferable doable across laboratories. The utility of the assay also makes it functional for non-contracted investigators to study chemicals and complex mixes that may not be under the purview of the Endocrine Disruptor Screening Program.

The validation of the Amphibian Metamorphosis Assay (AMA) involved three phases of validation. The first phase investigated how differences in exposure timing could impact outcomes and whether there was significant interlaboratory variation in outcomes. A multichemical study was also undertaken by the USEPA using both exposure timing scenarios. The second phase involved used the information derived from Phase I to formulate a standard operating document. This assay was then used compare exposure outcomes to several compounds with predicted thyroid or antithyroid activity across several labs. The third phase of the study evaluated a compound with strong endocrine activity (estradiol), but predicted not have direct thyroid hormone activity (please see comments below), and one with weak activity as evidenced in some literature. The validation studies demonstrate overall the utility of this assay for evaluating thyroid disrupting activity of the compounds tested. My comments below 1) address needed details in the final AMA Test Methodology, and 2) describe the limits of the assay as it was performed in the validation steps**.**

In reviewing the materials for the Amphibian Metamorphic Assay, I have followed the review Guidelines provided by the EPA. Some of my comments are general and not referenced to the page number on the Integrated Summary Report (ISR) and three Test Method Documents, and some are specifically referenced. Under section 8, I summarize my main criticisms of the AMA based on what issues that I evaluated under specific sections.

1. **Clarity of the stated purpose of the assay.**

ISR Pages 12-13: The overview and justification within the ISR is a brief review describing why the amphibian system provides a strong assay for investigation of the potential for anthropogenic compounds to impact thyroid related function. One addition that would be useful for this summery is a stronger overview of the timing of expression of thyroid hormone receptors during development compared with the release of thyroid hormone from the thyroid gland during amphibian metamorphosis. Such an explanation helps in the understanding of the set up of the two assay regimes that were tested in the Phase I validation

trials. Second, a brief overview of the receptors repressor verses activator activities might be useful ultimately for interpretation of outcomes, and because the receptors are expressed prior to increases in TH secretion. This information is critical to the understanding of the timing of the assay because the expression of TR during the earlier stages of the assay period (51-53) may lead to repression of TH sensitive genes and allow instead for growth of the tadpoles during this period, but if an environmental mimic is present, it could shift the activity of the receptor and accelerate metamorphosis. Buchholz *et al* (2006) is a useful review.

2. **Clarity, comprehensiveness and consistency of the data interpretation with the stated purpose of the assay**.

a. The endpoints are clear, not difficult to monitor, and appear to provide fairly consistent results across the validations. Some specific comments are below, however, regarding the interpretation of the data.

b. In the ISR, three possible outcomes are delineated on Page 22 Section 3.6 under Data Interpretation: "Thyroid Active, Thyroid Inactive, and toxic." The problem with this wording it that "thyroid active" does not distinguish between whether a compound is acting like thyroid hormone or inhibiting thyroid hormone activity. A possible change would be to have 4 categories (breaking up the first category in the original document into two categories that represent the different forms of "thyroid activity"). One possible suggestion would be "Thyroid mimic" and "Anti-thyroid Activity."

c. Sensitivity section page 49 ISR: The section in the ISR that tries to summarize which assay is most sensitive (14 verses 21 day) is not that clear. Although after several passes through the table, I was able to come to the same conclusion as the ISR, a brief summary table for sensitivity would be useful.

d. After Phase I, the decision was made to use flow-through systems not only for the other phases of the study, but also in the final AMA Test Method. However, no justification is provided for deciding to use the flow-through system. In other words, no statistical comparison was made to determine that this provides the better means of getting a more sensitive result (see further comments below).

e. A much stronger guideline for data interpretation within the AMA Test Method Documents is necessary. This issue was brought out when evaluating the Phase III estradiol results. In this trial, there lack of consistency in interpreting the estradiol exposure results when compared with the interpretation from phase I and II trial results. For example, the Phase II summary Table 6.1 in the ISR, Table X says there is no developmental effect, and then the report goes on to state that there is a significant reduction in the number of tadpoles reaching stage 60 in the estradiol groups. Is there an effect or not? This result suggests that investigators also should determine the number of animals not reaching a specific stage when conducting the AMA methodology. What was the statistical evaluation on the developmental stage across all groups? In the phase III study, clearly, more animals reached stage 60 in the controls than in the higher E2 doses (this finding is supported in a couple of papers in the literature in *Xenopus (eg.Gray & Janssens 1990))*, suggesting minimally, that E2 is interfering with thyroid hormone activity although the mechanism is not well understood. Also inconsistent is the fact that also found was a decrease in hind limb length which in the Phase III trial is considered to be general toxicity, but in the other toxicity measures are considered to be negative for toxicity. For example, these same findings in phase I and II would have lead to an interpretation of thyroid hormone antagonistic activity of E2. Such interpretation suggests that the data were evaluated based on the expected result for estradiol not being a "thyroid active" compound rather than on the outcome of the data. Last, there is a strong literature on the interaction between thyroid hormone and estradiol in mammals (Pfaff *et al.* 2000) and the receptors interact in ways that are complex (Vasudevan *et al.* 2002). This information suggests that, in fact, the Phase III trial demonstrated that estradiol may have thyroid disrupting activity. The Phase III results are very consistent with that literature and should be reinterpreted both to be consistent with the phase I and II studies and in light of

this literature. This issue of data interpretation comes up again in the Phase III BP-2 studies which also suggest that the effects of BP-2 on thyroid hormone function could be confounded by its direct actions and indirect actions because it also may act as an estrogen. And last, to further support this inconsistency in interpretation, in the IOP experiment of Phase III, lab 2 had the exact same findings (including decreased developmental stage and no thyroid histopathology impact) and these data are interpreted to be "thyroid active." *In summary, this phase trial demonstrated that data interpretation across the validation studies needs to be consistent, and guidelines need to be carefully developed to facilitate this interpretation.* In fact, in the AMA Test Method, there is no section on data interpretation, and in the overall ISR, there are no clear guidelines for how many parameters need to be significantly different from controls before a compound is to be interpreted as thyroid disrupting. Such guidelines are essential and should be provided clearly in the final AMA Test Method Protocol, along with appropriate summary tables.

3. Biological and toxicological relevance of the assay as related to its stated purpose.

a. In the validation processes it would have been useful to determine whether an exposure protocol from hatching through metamorphosis provided a different outcome than either of the shorter protocols (stage 54/14 day or stage 51/21 day). One of the issues this assay did not answer in the context of the presented validation phases is whether early exposure (prior to stage 51) impacts later thyroid-related outcomes. Animal (including human) populations may be exposed to these compounds throughout their lives, not just from a specific stage on. This early exposure may have impacts on the thyroid system that will not be seen here. For example, we have preliminary data in our lab where we see impacts on timing to metamorphosis from exposure to complex mixes that we do not see with a 14 or 21 day assay (Propper, unpublished data).

b. In this assay validation approach, through 3 phases of the validation process only three known environmental contaminants were evaluated. These are perchlorate, which has well-documented means of thyroid hormone disruption, estradiol which in fact does have some affects on the thyroid hormone system (see comments above), and benzophenone-2. The other compounds tested were chosen largely based on their known pharmacological thyroid hormone disrupting activities.
A final validation step evaluating some environmental contaminants (eg. Pesticides or pharmaceuticals) known to have thyroid disrupting capacity at environmentally relevant concentrations is necessary for final knowledge of the utility of the AMA Test Methodology. For example, endosulfan has demonstrated impacts on thyroid gland histopathology, and impacts thyroid hormone levels in a number of species (including human pesticide formulators). A final validation step demonstrating the usefulness of this assay using a common environmental micropollutant would strengthen the justification of the protocol. Another advantage of including such a validation is that the probability of getting a monotonic dose response is much less, and therefore tests the validity of the assay when a U-shaped (or other non-linear response) is generated.

4. **Clarity and conciseness of the test method in describing the methodology of the assay such that a laboratory can:**

a. comprehend the objective,

The objective as stated in the test methodology is very short and to the point; however, it would be useful to provide references or weblinks to the other documents that were provided to the peer reviewers so that the labs conducting the tests have access to all the justification for the development of the assay**.**

b. conduct the assay,

a. In general, the test method is missing several details that are necessary. First, there was interlaboratory variation in the validation phases of the test methodology development. To minimize such variation, the

assay methodology must be very clear and detailed with acceptable alternatives to the specific methodology clearly delineated (as well as unacceptable alternatives). Such detail is necessary to insure 1) that there is consistency in approach among any EPA contracted laboratories, and 2) that there is consistency in use of the assay by non-EPA researchers who are trying to adopt this assay to their labs' specific hypotheses. Specifics are addressed in the context of the specific heading within the AMA Test Method document.

b. Exposure System: The exposure methodology needs more details in several areas outlined below.

1. The flow-through system is designated as the system of choice, but an option is provided for static renewal. There are problems associated with the justification of the choice and with the description of the methods for using either of the choices.

a. Static renewal: If static renewal is to be used then details of how the water is removed or how the animals are to be transferred to clean tanks needs to be carefully addressed. Then if tanks are to be reused, methods for cleaning and rinsing all the glassware between water changes also need to be provided.

The AMA Test Method protocol states that a complete water change is made if the static renewal system is to be employed. This method implies that the animals must be moved which can cause stress and damage to the animals. A complete water change also removes any bacterial communities that have developed in the tanks that may be necessary for appropriate tadpole development (although if complete water changes were used in the German lab, then it may not represent a problem as the controls performed similarly to the other labs using the flow through system).

I have searched all of the provided documents, including the methods for the Phase I trail and in the "Annex" of the Phase I trail report for the details of the German lab's methodology for static renewal. There is no description of how the static renewal was conducted (and in fact in the "annex," the method is referred to as "semi-static." What does "semi-static" mean?). The provided AMA Test Method provides very few details except that there needs to be a complete water change at least once every 72 hours, and every 24 hours if justified by criteria that are not well defined in the document. A whole water change every 24 hours will be extremely stressful for the animals, and since stress and thyroid interact in this species to impact developmental timing, such frequent water changes must be avoided. If contracted labs are allowed to use the static renewal approach, much more detail needs to be provided in Final AMA Test Method document, including handling of the animals between water changes, whether the entire volume of the water is changed, and how the animals are to be dosed. Also, whether all the replicates are to be refilled from a common water source with the exposure chemical diluted, or is each replicate dosed independently, needs to be considered.

b. Flow-through system: Again, more details need to be provided. First, does each replicate tank receive an independent water source made up by independent dilutions of the stock solution, or do they come from a common water source (I recommend the former to maintain replicate independence). Second, one type of plastic tubing is recommended in the AMA Draft, but the method states that other unspecified types are acceptable. It is absolutely critical that both acceptable and *unacceptable* plastic tubing be listed. The method needs to specify that the supply tanks must also be glass, and how often the supply tanks are refilled needs to be specified as well. For example, should the tank be refilled daily, every other day (clearly a larger volume will need to be made from stock), or weekly? For pumping the water from the supply tanks to the exposure tanks, more detail would be helpful. Getting exactly 25 mls/min via gravity feed is not easy, and making sure each tank gets exactly the same flow rate would be very difficult indeed. Inexpensive pumps that can be set for such a flow rate should be recommended.

2. Adult Care and Breeding: Consistency in the breeding protocol needs to be strong, and the detailed methodology should be provided here and not just referred to an unreferenced FETAX methodology. Also, it needs to be made clear that using older frogs can lead to delayed development in the tadpoles. The

breeding frogs should be purchased for breeding not more than a year before the study. This information is buried deep in *Xenopus* breeding information available online, but I have personal lab experience to attest to the fact that older animals produce slower developing larvae.

3. Larval Care and Selection pages 5-6 AMA Attachment A1. This section needs much more detail.

a. Using tadpoles from one spawn is insufficient. If animals from only one pair of breeding animals is used, any effects (or lack of effect) from exposure found may be strictly due to the sensitivity (or lack of sensitivity) of the one pair's offspring. Three spawns from three separate breeding pairs are really the ideal. Equal numbers of animals from each spawn can be distributed among the tanks. It may increase the variation slightly, but it avoids the risk of pseudoreplication based on a sample size of 1 spawn. In the mammalian literature, peer-review would never accept data supplied from the treatment of 1 litter alone.

b. Is the 2% cysteine placed in the breeding media or in the culture media?

c. *What is the culture media for raising the hatchlings and what is the culture media for rearing during the exposure? This detail is critically important*. In the Phase trials there were some differences among controls suggesting that the media may be important. For consistency, one type of control media should be recommended and made up from preferably deionized or even e-pure water that has the salts (including iodide) added back. Experience from my lab precludes dechlorinated charcoal filtered tap water (there is still something in that water that is toxic to our animals). Other labs may find similar problems. There are several potential options that would lead to consistency in growth media. Labs should either use FETAX (very unpopular among some researchers I have communicated with, but still used by others), 10% Holtfreter's media or some other modified water with salts added back (some *Xenopus* supply outlets even provide their own salts), *but <u>one</u> version should be chosen for the AMA Test Methodology, and it should not be region-specific tap water.*

If the culture water for rearing is the same as for exposure, it needs to be explicitly stated. If it is to change, for example from FETAX to some other media, that needs to be noted, and again, one type of water (not regional tap water) needs to be chosen. Also, once exposure starts, should the exposure tanks receive the water for a specific amount of time before transferring the tadpoles to the tanks? Last, I would recommend that the tanks must all be aerated during the exposure period and that the DO is measured daily in all tanks and noted.

d. What is the density of the animals in the hatching tanks? What is the volume of the tanks, what is the volume of the culture media in the hatching tanks? All of this methodology should be provided.

e. Are the clutches from each spawn mixed in the hatching tanks (they should be, but if not, they need to be evenly divided within each replicate for all treatments: see comment 3a above)?

f. Under "Larval care and selection," the Table 2 on page 6 should be clearly referenced.

g. The Pre-exposure protocol, page 5 needs more detail. If this pre-exposure period is supposed to provide conditions similar to those of the exposure period, then 1) Static renewal should only be used if it is to be used in the exposure system to, and 2) the flow-through rate should be the same as in the exposure period (25ml/min).

h. Is the water volume reduced once the 5 tadpoles are removed on day 7?

4. Dosing:
<u>a. Analytical Chemical Sampling page 6 AMA Attachment A1:</u> It needs to be made very clear that the quantification of the exposure chemical is to be done for each replicate not just for a representative

replicate. In the flow-through system, it also should be made clear that the supply tanks be measured at least once at the beginning and once at the end of the experiment. Details for how much water is to be removed or for determination of such for the chemical quantification needs to be supplied. Further, how the samples are to be stored needs to be provided. It may be that for new compounds such information is limited, but guidelines need to be developed and provided for this methodology.

b. Dose Determination page 7: The issue of dosing is very complicated, and the basis for the decision making outcome is not adequately addressed in the AMA Test Method or in the other documents. The decision to start at a dose that is at the maximum tolerance level (10% mortality) or 100 mg/L, whichever is lowest, has little justification based on the endocrine disruption literature. This level can be at the 100 parts per million range which can be anywhere from 10,000 to 1,000,000 fold greater than is often seen for endocrine disruption. Furthermore, such dosing would potentially lead to compounds being tested at ranges that would far exceed their levels in the environment. Given that: first, many studies have shown thyroid disrupting effects at levels well below these recommended exposure levels; second, the impacts on the endocrine system often do not show a clear linear dose response; and third, this level of testing does not take into account the potential levels of the compounds in water or sediment, how will the results be interpreted in a regulatory environment given that no effect level may not be found with the minimum exposure dose being potentially 11 mg/L?

5. Attachment A2: Embedding tissues. There is one inconsistency: Part 9. States that the head is oriented either ventral to dorsal, ventral side down or "rostral to caudal" and then "caudal side down." To be consistent, need to state that the head is oriented "caudal to rostral" caudal side being the leading edge of the block.

6. Attachment A2: Sectioning tissues: Part 4J, page 9. This section is critical and therefore needs more detail. It would help to state at the beginning how many final sections are to be mounted and stained, and about where in the tissues these sections are to be collected. Having done a lot of histology, it is possible for me to take a best estimate of what is suggested by this methodology, but the step sectioning and examination of the sections prior deciding which to finally mount is not written very clearly.

c. observe and measure prescribed endpoints,

1. Why is 10% chosen as an acceptable mortality rate when in the Phase Trials, 5% was the maximum acceptable mortality rate? No justification is given for this shift or for the 10% rate within the explanation of the test methodology.

2. Under determination of Biological Endpoints AMA Attachment 1 Test Method, beginning on page 8:

a. A URL link or reference to *Xenopus* staging with pictures should be provided within the test protocol.

b. Additional Observations (page 10):

ii. Behavioral Observations: If behavioral parameters such as uncoordinated swimming, hyperventilation, quiescence, etc are to be "observed," they need to be done so in a coordinated and quantifiable fashion. One methodology would be to do a 1 min focal animal observation on 3 animals/tank/ at day 7, 14, and 21. In the current protocol, there is no standardized way for making these observations and analyzing the results.

ii. Grossly Visiable Malformations: A list with pictures, if possible, of the usual gross morphological problems needs to be included in the protocol (kinked tails, bent backs, extra limbs). These problematic

gross morphological outcomes should be included in the final evaluation at 7 days and 21 days and should have their own column in the data spreadsheets.

c. Under Test Initiation and Conduct: Day 7 (page 10):

If thyroid histology is to be conducted on Day 7, then it needs to be clearly stated here. If not, then there still needs to be a statement saying this subsample of the animals are to be stored individually in Davidson's Fix and then 10% NBF.

d. Under Data Collection and Reporting (page 12): Overall, the data tables supplied are adequate, but some additions, especially in summary tables would be helpful. Also, supplying a Quality Assurance Plan is necessary. It is referred to here, but not provided in any documentation.

Under Chemical Observations and data (page 12): Details need to be provided for how to collect the water for these determinations. Instrumentation should be identified, and SOPs provided as an appendix for everything except temperature and pH. This protocol should facilitate the ability of contracted and non-contracted labs to conduct an assay as similar to each other as possible. Also, if actual measures of test chemicals in the water are to be taken, then why might stocks also need to be measured? The way the protocol is worded now is very vague (states, "may be required") about whether the stocks need to be tested.

3. Attachment A2: The title needs to change so that the morphometric measurements come into play in the first part of the title. The title as reads emphasizes the histopathology. One suggestion is "Guidance Document on AMA Endpoint Sampling Part One: Technical Guidance for Morphological Sampling and Histological Preparation."

4. Attachment A2: Trimming of tissues. More detail is needed for how to remove the mandible for histological preparation. No detail is provided here.

5. Attachment A2: Image analysis. For each parameter that is digitally quantified, at least 2 measures should be taken and then averaged as there is some variation in how the lines are drawn. Also, one person should conduct all the measures across all treatment groups and should probably be blind to the treatment when conducting the measures.

6. Attachment A3: Some of the measurements of thyroid gland histology could be done via direct image analysis and direct quantification rather than semi-quantitatively by grade. However, this process is laborious and time consuming. The grading scheme, with proper training, and good preparation appears to be justified, and appeared to work for the assay in the validation data presented.

7. A Quality Assurance Plan document is mentioned, but none was provided. It should be an attachment or appendix with the AMA Test Method.

d. compile and prepare data for statistical analyses.

Under Statistical Analysis:

1. Mortality data cannot be analyzed by an Anova. Some form of G-test or Chi-squared will need to be employed.

2. A Mann-Whitney is employed if there are two treatments. For more than two treatments, first a Kruskal-Wallis should be employed first, followed by Mann-Whitney. Also, there is confusion across

documents about how to conduct the statistics if the treatment effects are a linear verses non-linear dose response. Since in these types of studies (at least at low dose exposure) non-linear effects are often found, what type of statistic should be employed?

3. There is some confusion in the Phase trials about whether HLL should be standardized by body length or not. In the final data tables in the AMA Draft Test Methods, there is no mention of whether or not the HLL should be standardized. It needs to be made clear, prior to doing statistics, about whether this parameter should be standardized for final analysis and interpretation or not.

4. The Fig. 8.1 flow chart in the ISR has thyroid histology following only negative results in other areas. However, in reality, each of the tests conducted histopathology. If the EPA wants histopathology conducted always, then this assay needs to be placed higher up on the logical flow for data interpretation chart.

e. report results.

1. Many data sheets are provided, but no guidelines for data interpretation are provided. In the three phase trails used in the validations, there were decisions made regarding the outcome interpretations, but in the test method, there are no guidelines. Once the data are collected and analyzed, how will they be interpreted? Summary tables like the ones used in Tables 4.5, 4.8, 4.12, etc. in the ISR should be provided to facilitate overall interpretation of the data. In these tables, instead of individual labs being columns, the dose of the compound used could be used across the top of the table. In fact, in the phased validation studies, one of the criticisms I have is that there was no information in the data summary tables of the doses considered to have effects. Again, this problem can be addressed in a final summary table provided in the AMA Test Method that allows for data interpretation across doses. Such a reporting system will also help in interpreting the data if non-linear effects are seen (see comments above).

2. Throughout the phase trials and in the Draft Methodology, there is no mention of how final decisions are to be made regarding the outcome of the test (mentioned also elsewhere in this review). It may be possible to combine all parameters measure and to apply a principle component analysis to determine the outcome of the exposure. Alternatively, consistent approaches to the data interpretation can be developed, and followed carefully. No matter the approach, it needs to be carefully outlined in the final Test Method.

If warranted, please also make suggestions or recommendations for test method improvement.

I have incorporated most of my suggestions in my comments above, and I will summarize the main points below under section 8.

As asynchronous development is an important endpoint as pointed out a least twice in the summary documents, it will be critical to provide labs with a clear-cut standard operating procedure for scoring this issue and analyzing and interpreting the data.

5. Strengths and/or limitations of the assay.

a. The strength of this assay is in its ability to determine whole animal disruption of thyroid hormone-related physiology. One weakness is that the assay itself will not determine how the disruption is occurring.

b. One limitation of the assay is that animals are not dosed throughout development. Such testing may lead to increased sensitivity of the assay (see comments above).

c. A major limit to this method is the number and choice of doses used in the assay. Little justification for the dosing decision-making process is given. The doses are decided based on the overtly toxic dose. The ISR needs to present a clear rationale for this dosing approach, and it needs to be made in light of the literature in the field. The decision to only go with three potentially very high doses that do not even differ by even 10 fold is a mistake. First, the literature in endocrine disruption demonstrates time and again that there are non-linear responses, especially at very low doses. Second, environmental exposures to many chemicals in the environment are occurring at the part-per-billion or even part-per-trillion levels. The current dosing regime for this assay would most likely be well above these levels. Last, there is the issue of non-linearity of response, especially at the lower doses that are important given the risk of exposure to human and wildlife populations are mostly at low doses. In summary due to the non-linearity of some dose responses and the fact that a very low dose can have more impacts on endpoints than higher doses, these doses need to be evaluated. The AMA should be sensitive enough to pick up on these low dose and non-linear responses.

d. Because of the issue of non-linearity, this methodology needs further development with how to deal with no- linear dose responses. The report was unable to really respond to the occasional non linear response, yet in many endocrine disruption studies, the finding of non-linearity is the case. A clear approach is necessary. For instance the final scientific review panel may state that if any dose has an effect, the result is a positive. Alternatively, they could decide that if two of the 3-4 doses tested need to be positive before they determine an effect. How will these types of non-linear results be interpreted?

e. The methodology (and in fact the validation trials) do not provide much information for reporting the dose effects. The overall reporting is a yes or no outcome in the reporting tables for the phase trials with no information provided about the lowest effective dose level. Dose effects need to be taken into account in the final reporting for the assay.

f. One last limitation is the lack of how this assay can address the issue of exposure to complex mixes. The field of ecotoxicology is still in its infancy regarding evaluation of the complex mixes which are what all organisms are really exposed to. Furthermore, mixes can interact with each other to lead to endpoints that individual compounds will not. Even thyroid hormone and estradiol interact (see comments elsewhere in this peer-review). Can this AMA protocol be applied to testing for understanding the thyroid hormone disrupting capacity of complex mixes? Even if the EDSP purpose is not to test mixes, others in the field will want to adapt these protocols as closely as possible to their studies.

6. Impacts of the choice of test substances and methods chosen to demonstrate the performance of the assay.

a. The choice of the tadpole metamorphosis system as a test assay is outstanding given the knowledge base of the system, and the relative ease of use and data interpretation. The comments below are to the details of the method and not to the overall utility of the assay. Once the methods are standardized and clearly detailed, this assay will undoubtedly provide a useful measure for thyroid hormone disruption.

b. In the development of the assays, one lab used static renewal methodology while the others used flow through systems. Ultimately, the ISR states and the AMA Test Method Draft recommends the flow through system with little or no justification based on the studies. The data clearly demonstrate no difference between the two systems in control performance. Also, in the validation no data are provided for the actual dose received in the static renewal system. If static renewal is to be allowed, it is critical to know if the concentrations of chemical treatment (dose the animals receive) are equivalent between the two systems.

c. As mentioned elsewhere in the review, the assay is validated using compounds that are known agonists or antagonists of thyroid physiology. Also, a presumed non-thyroid disruptor was evaluated (estradiol; see problems with data interpretation above) along with a weak distruptor (BP-2) at fairly high doses. It would be useful to have one more validation step using a pesticide of some form that has known thyroid hormone disrupting effects at environmentally relevant levels.

d. The interpretation of results with a compound like IOP is very interesting, and needs to be carefully evaluated, as some of the compounds likely to be tested via the EDSP may have such complicated modes of action. The results on HLL may be difficult to interpret given the impacts of such compounds also on body length, but then it is possible to do the analysis as an index: HLL/BL.

7. Repeatability and reproducibility of the results obtained with the assay, considering the variability inherent in the biological and chemical test methods.

Overall, the interlab variability was minimal, however, there was some variation and testing may need to be conducted independently in at least two separate labs.

8. Please comment on the overall utility of the assay as a screening tool, to be used by the EPA, to identify chemicals that have the potential to interact with the endocrine system sufficiently to warrant further testing.

Overall, the AMA will be a useful screening tool for testing compounds and complex mixes for thyroid hormone disruption. There are some details that need to be added or clarified within the assay protocol itself, and some addition information/validation that might prove useful. These issues (all brought out above) are summarized below:

Summary Points:
1. The outcome of "Thyroid Active" needs to be divided into two categories to provide information regarding whether a compound has agonist-like or antagonist-like activity.
2. More detail is necessary in the set up of the assay and in the delivery of compounds.
3. Clear consistent control water from DI or e-pure water with salts and iodide added back must be used rather than region-specific tap water.
4. More detail is needed in the raising of tadpoles to stage 51, and in what type of water should be used for the first days of growth, and at what stage to switch to the water for culturing the animals during the study period.
5. Tadpoles from multiple spawns should be divided among all tanks and treatments.
6. There is a strong need for clear guidelines for data interpretation. The phase trials provide tables with a thyroid disruption +/- scheme, but the interpretation of the presented results across the three trials are not consistent. For example, would the testing lab conclude that a compound is thyroid disrupting if at least 2 criteria are met? How about 3? In other words, how will those summary tables be used to determine whether a compound has thyroid-like activity, blocks thyroid hormone function, is not thyroid active or is toxic. Again, there was inconsistency in data interpretation across the Phase trials.
7. Dosing needs to be over a wider range and needs to have some treatments that are within predicted exposure levels for human/wildlife populations (low ppb range).
8. Mechanisms for reporting dose outcomes and overall dose limits of sensitivity need to be developed.
8. A final validation step needs to be undertaken to evaluate one or two more compounds known to impact thyroid hormone function. These studies should compare the outcomes of the doses determined as described in the AMA Draft Test Method to environmentally relevant doses.
9. Concern exists for the interactions of these compounds. One of the main limits of any of the EDSP assays is that they do not address the impacts of complex mixes of compounds. No organism is exposed to any one compound, and it needs to be noted in the final version of these assays that a negative finding for

the potential for endocrine disruption cannot preclude that the compound might interact with others to have endocrine-relevant impacts.

Reference List

Buchholz DR, Paul BD, Fu L & Shi YB 2006 Molecular and developmental analyses of thyroid hormone receptor function in Xenopus laevis, the African clawed frog. *Gen.Comp Endocrinol.* **145** 1-19.

Gray KM & Janssens PA 1990 Gonadal hormones inhibit the induction of metamorphosis by thyroid hormones in Xenopus laevis tadpoles in vivo, but not in vitro. *Gen.Comp Endocrinol.* **77** 202-211.

Pfaff DW, Vasudevan N, Kia HK, Zhu YS, Chan J, Garey J, Morgan M & Ogawa S 2000 Estrogens, brain and behavior: studies in fundamental neurobiology and observations related to women's health. *J Steroid Biochem.Mol.Biol.* **74** 365-373.

Vasudevan N, Ogawa S & Pfaff D 2002 Estrogen and thyroid hormone receptor interactions: physiological flexibility by molecular specificity. *Physiol Rev.* **82** 923-944.

Hannes van Wyk Review Comments

1. **Clarity of the stated purpose of the assay**

The Introduction and stated purpose of a Tier 1 assay was clear. Personally I think the general explanation of the purpose of a Tier 1 assay is extremely important. I don't think it is always appreciated what the actual purpose of a Tier 1 assay is. In the Introduction and background to the stated purpose of the assay the progression of assay development, validation and evaluation are important components to the reader. In order to underline the role/place of a Tier 1 screening assay in the larger picture of assessing EDC activity I would like to see a diagramme showing the contribution of Tier 1 screens. The criteria set by EDSTAC for Tier 1 screens were presented. With this statement "*It is important to recognize that the AMA is not intended to quantify or to confirm endocrine disruption, or to provide a quantitative assessment of risk, but only to provide suggestive evidence that thyroid regulated processes may be sufficiently perturbed to warrant more definitive testing*" the purpose of the AMA is placed within the framework of a Tier 1 screening programme underlining the purpose of such an assay and sets the scene to understand the development and validation of a Tier 1 assay.

2. **Clarity, comprehensiveness and consistency of the <u>data interpretation</u> with the stated purpose of the assay**

Data interpretation in some cases was difficult. It is acknowledged that in terms of the size related endpoints non-thyroidal effects may have been operational. Throughout the authors tried to do comprehensive interpretations and were mostly consistent focusing on the fact the AMA is a screen for thyroid interaction. They must be credited for constantly viewing the methodology, set-up and experimental design for possible explanations for inconsistencies. They made some effort to understand lack of reproducibility among laboratories. Interlaboratory validation data were presented in a logic manner, making the assessment easy. In the final interpretation an improved logic interpretation progression was proposed and the summary data presented according to this proposal. This was helpful since the opportunity to tests the proposed scheme was used effectively. In some of the inter-laboratory data-sets, one got the feeling in spite of inconsistencies or low reproducibility/ repeatability the final conclusion was clear-cut. So it was difficult to comprehend what the threshold (within the weight of

evidence perspective) was for making the particular decision. But, overall seen, the data interpretation was sufficiently clear.

3. Biological and toxicological relevance of the assay as related to its purpose

In all the literature presented, including earlier DRP and recent published literature, toxicological relevance of assays focusing on environmental (external) thyroid modulation with potential adverse consequences for wildlife and human health, will always be relevant. An extensive literature now exists that suggest a range of environmental toxicants that may in some way interact with the thyroid system. The Introduction and Background sections of all the documents recognize this phenomenon.

As acknowledged by most authors, the general control and endpoint expression associated with the thyroid system is rather complex. More the reason to understand the range of potential mode of actions to ensure toxicological relevance. I am convinced that this point is clearly made in the "Rationale for the assay". The rationale for employing non-mammalian organisms as models for assessing thyroid disruption seems to be convincing and acceptable, especially when considering the recognized evolutionary conservatism among vertebrate groups. The AMA uses the advantages that amphibians offer to study endocrine disruption of the thyroid system through phenotypic thyroid hormone (TH) dependent changes during the developmental phase (metamorphosis). The role of TH during early amphibian development (with free-living embryos) and early mammalian development underlines the relevance of using the AMA, a simpler more straightforward system to work with than working with early mammalian life stages.

The authors clearly and comprehensively reasoned the relevance and advantages of using an anuran metamorphosis model in studying external influences on the thyroid axis. In Section 2.2, they summarize the dynamics of hormonal changes during the developmental programme. Similar changes in expression of TH-receptors were presented elsewhere. It is clear that during the development, refinement and validation phases of the AMA considerable thought has been given to the relevance of the exposure window. It is also clear that several possibilities exist to use short term, molecular based TH receptor expression along with the longer-type assay using morphological based endpoints. Although, it seems that earlier suggestion for the inclusion of the former did not materialize as integral part of the AMA.

The importance of controlling for several environmental conditions that may secondarily affect the rate of metamorphosis was also shown. This must be valuable to the user of the AMA, specifically to understand the sensitivity of amphibian development to a range of environmental factors and therefore the importance of controlling for these to ensure the correct interpretation of exposure data.

The authors adequately describe the possible points of modulation and uses Figure 2.2 to show the non-neuro-endocrine (or peripheral) points of concern. It is not clear why they selected to omit the potential points of effect on the neuro-endocrine side?

Reading the DRP and the ISR together, I am convinced that the extent of literature review to set the scene and build the rationale for the AMA is extensive and represents a good review of the literature to highlight the hormonal control of amphibian metamorphosis. It has been shown that the AMA represent an opportunity to study several TH dependent endpoints and mode-of-actions rather than just screening for the ability of chemical to bind to TH receptors (like in several HTPS assays). Apart from the classical genomic interactions, non-genomic interactions as well as pathway enzymes involved in synthesis and metabolism activities may also be included. In summary, therefore I am convinced that the biological and toxicological relevance of the AMA has been shown. Although it runs the risk of being too "reductionistic" when it comes to EDC action, it represents a broader multi-endpoint perspective, and therefore, certainly conforms to the goals of a screening assay for suspected/potential Tier 1 EDC interaction.

4. **Clarity and conciseness of the test method in describing the methodology of the assay such that a laboratory can:**

a. Comprehend the objective

The AMA is structured in such a way that the laboratory should be able to comprehend the objective of the tests to eventually answer the questions related to the purpose of the assay. The selection of *Xenopus laevis* as the test species is explained in the ISR as well as in the DRP. In the DRP comparisons are made between potential test species. From all this it seems that *X. laevis* is still the appropriate species to choose. One aspect of concern is the fact that hCG is used to initiated breeding in captive populations. Very little information on the potential effects of hCG on the response of the thyroid axis to external compounds are available. This is especially concerning when considering the dose of hCG used. Although the AMA will be used to screen chemical compounds and hopefully also mixtures of compounds, therefore, in laboratory studies, the use of local endemic species will have the added advantage of answering environmental questions. However the fact that *X. laevis* is fully aquatic makes the exposure protocol simple. Table 5.2 seems to be a good summary of comparisons among different candidate species. (The reference to *X. tropicalis* as a South African clawed frog is incorrect, West African?). In summary, enough evidence are available that suggest that *X. laevis* is a robust model and currently the best amphibian species available suited for use in the AMA, with several advantages in handling and breeding of tadpoles for in-laboratory exposures. However it may well be that several other amphibian species could also be used to answer specific questions regarding thyroid endpoints. The knowledge explosion regarding *X. laevis* clearly makes it a valuable aquatic indicator species. Models, like *X. tropicalis* and other local endemic species, may in future be used to answer specific questions, but in the mean time *X laevis* seems to be the best studied non-mammalian model to study aspects of thyroid functionality.

b. Conduct the assay

Breeding of Tadpoles: As mentioned before I have a concern about the use of hCG in general but secondly the dose applied seem rather high. Successful breeding and tadpole production can be obtained with much lower concentrations. Although the higher dose ensures large number of tadpoles, the question of secondary effects comes into play. The question of seasonality may be a problem if the laboratory received recently collected frogs from South Africa. Using frogs collected from natural sources for breeding purposes show some seasonality in terms of response to hCG stimulation and egg production. Whether this response is lost with acclimatization and after what period of acclimatization is not known to me.

Following spawning, the SOP states that that the best spawns should be retained. This decision is based on embryo viability. How is this determined? Hatched embryos should be removes as soon after hatching a possible since the water quality goes bad soon after hatching because of all the unhatched eggs. Not convince that the cysteine treatment is necessary. Also not sure about the pipet collecting method. The suction action of the pipet may impact on the embryo. Netting free swimming hatchlings with a flat scoop net seems better. Density control is important during development.

Staging of tadpoles: Although Nieuwkoop & Faber (NF) staging is not too difficult, the criteria used to stage the tadpoles are not clearly stated. I feel more effort should be made to describe the characters to be used (or show visually). Size (WBL) may be variable. N&F state that the optimal size at NF stage 51 is 28-36mm but the Appendix A1 give a range of 24-28mm. NF stage 51 describes the forelimb as oval vs conical in Stage 52, the hindlimb as conicle in shape and the length of the hindlimb as 1.5X its breadth. For a newcomer the staging may be difficult and more detailed or clearer description of the important stages are needed, in particular for the landmark traits. A table summarizing these landmark traits with a pictorial guide will ensure more accurate

staging. Stages 51-57 are based on the growth of the hind- and fore-limbs. Stage 58 states that the forelimb is free from the atrium (a landmark). Then criteria switch to aspects of the forelimb (length to hindlimb). Is this how EPA is using the staging? N&F include detailed descriptions of all organ development. The question is which of these can be used confidently by persons doing the staging? Standardization of criteria used, otherwise more variation

Selection of exposure period and length of exposure: The selection of the exposure window did get some consideration during the refinement stage of the assay. Clearly this aspect got considerable thought, discussion and testing in the end. It therefore seems acceptable to use the assay in the suggested window for a period of 21 days.

Exposure procedures (set-up):
Very limited information or reference to the protocols suggested to dissolve chemical in treatment water was given, in particular the liquid-liquid and glass wool saturation systems. If carrier controls are included?

Flow-through vs semi-static. Although in the initial descriptions of the exposure system, the semi-static renewal systems was described in detail limited information is given about the design of the flow-through system. Flow should be low since in its natural environment, *X. laevis* tadpoles occur in low-flow situations. It has been mentioned in Appendix A1 that if semi-static exposure is used, the concentration of test chemicals should be reported and that a 24hr renewal interval is ideal. The question is how practical and cost effective this is to measure the chemical concentrations in the water samples. Did the authors mean that in a preliminary study the dynamics of the mother compound in the water column must first be established?

Exposure procedures (control chemicals):
The experimental design seems adequate. However in a flow-through system the question arises regarding the effluent produced and how it should be handled/discarded. Although several types of flow-through systems are potentially available, the authors don't give enough information on the diluter and flow-through system they used. They also don't describe and discuss the options of different semi-static systems that may be used or have been used by the German laboratory. Detailed SOP for using these systems are lacking and a laboratory that hope to do the AMA will find them in a vacuum.

c. Observe and measure prescribed endpoints

Developmental stage:-I am not convinced that all labs will extract the same criteria from N&F (1956) to determine the stage. Some guideline must be given and I feel detailed description of criteria used to stage a tadpole is necessary. The N&F (1956) document is not very friendly to read. How will one handle asynchronous development, making it difficult to stage a particular tadpole, using the standard trait set? Mention has been made of differential characters, advanced characters in the head region and arrested characters in the hind limbs. What set of characters are practical/important? The authors state that the staging is simple and clear-cut. I am not convinced about this.

Hind Limb length:- Gene expression studies show that the measurement of hind limb as endpoint make good sense. However, I am worried about not enough detail given as a SOP to measure hind limb in a standardized way. Especially when the limbs are long and well-developed, the line one takes when measuring may influence the outcome. Figure 1 in Appendix 1A is rather simplistic and does not show the real situation. The revised photo presented in the histological appendix does not solve this problem. Detailed landmarks are necessary for consistent results when applying a general bio-assay.

Body length and Weight:- In practice the opening of the vent is quite a difficult point to measure as well. Should one not use the base of the vent as an alternative measuring point/landmark?

Thyroid Gland Histology:- Numbers collected at day 7 and again at day 21 for histology will generate a large number of histological samples that need to be processed and eventually evaluated. The selection of individuals? 1) Why a day 7 sample? 2) The selection of samples seems rather complicated. 3) difficult to see it being practical to select randomly but also to try and stage match (later I see they actually recommended stage matching (see below). This will only be possible if one chemical is done at a time (with dilution replicates) and compared to a control. The absence of stages in treatment groups make stage matched comparisons difficult and will an extended control sample be necessary to generate same-batch control stages (see suggestion below).

d. Compile and prepare data for statistical analyses

Not clear enough. Maybe the use of a diagramme will aid the understanding of the data grouping and analyses.

e. Report results

The outcome seems clear. But, to come to a conclusion will take some interpretation, especially if the correlation between histological data and morphological data is weak.

Separating non-thyroidal toxicity from thyroidal effects will be problematic and criteria used vary vague. In particular, at the lower-end of agonistic and antagonistic effects.

Interpretation of Histopathology will have to be done by an experienced pathologist. Will it be possible to build this capacity in the laboratory or will expert scientists be contracted to do this part? How many amphibian pathologists do we have in the world, or will a human or wildlife pathologist equipped to do the screening?

More specific guidelines should be giver regarding the presentation of data. Can the reporting layout be made standard to ensure reporting of the data as well as assay performance data?

f. Test method improvements

More detailed SOPs are needed. The earlier suggestion by the German, French and Irish scientists (DPR) that a short-term gene expression study be included seems to make sense. The initial response that the level of complexity of this technique and the difficult interpretation of the data may not be valid since histological interpretation seems to be rather complex as well, although cheaper to produce. Investigating laboratories could out-source these aspects to specialists (will probably have to do it in the case of histopathology anyway).

Discussion of refinements suggested in the ISR:-

Dietary regime:- I agree that the feeding regime must be standardized or monitored in relation to a few growth performance checks, say at 7 days and 14 days in the control groups.

Water iodine levels:- I agree with this suggestion. The question is, has this problem been researched adequately?

Dose levels:-I agree with this suggestion (see below)

Stage matching for histopathology:- I agree with this refinement, however, the comparison is with the Control group and in some cases you will not have matching stages (either in antagonist or in agonist groups). Two possibilities may solve this problem: 1) if the performance criterium of Control stages being around NF 57 then comparisons could be made to known histology documented from all developmental stages (Atlas approach), but, if it is better to compare with internal control, then initial control sample (groups should be increased and representative stages sampled at certain times to facilitate stage matching with controls. Following 21 days, remaining Controls can be maintained to reach stages reached in agonist groups. At least for this batch stage matching will be possible. This problem only occurs when working with strong antagonistic and agonistic chemicals.

Improved Data interpretations:- Agree with the suggestion, but would like to know why only use "advanced development" to get a "Yes" and therefore exclude the need to do histology? Why not also include "advanced inhibition/retardation" to get a"Yes"? I know there was a concern that the histopathological assessment is time- and specialist-consuming and should therefore be limited. But, without direct evidence of some kind, for example, molecular (thyroid receptor (TR) expression)) or histological evidence the risk of getting a false positive (agonistic or antagonistic) seems to be greater? Refer to Table 8-3 for T4. If I understand correctly all labs will conclude "active" after noting advanced development. However, the histology only supported this conclusion in two of the labs. I can see that the compound will still move to Tier 2 but at least more direct knowledge will be available regarding the histopathological picture.

Another suggestion: I am of the opinion that collecting of material (in RNAlater for example) during the exposure phase (either independently at 48 hours or after 7 days) could add another level to Figure 8-1. QPCR technology is becoming more and more routine now and could greatly aid as a last step just to make sure you don't have false negatives.

Strengths and limitations of the assay:- I agree with the discussion on potential limitations listed, but also underline that several of these represent knowledge gaps. The use of non-mammalian models as early warning systems to human health still has to go a long way. However, the appreciation of interaction between environment and organism will flow from such aquatic non-mammalian models. While it is true (point 2) that morphological and/or molecular responses may be different in developing young and adults, the effects at the developmental level by several EDCs are the most dramatic, both at short-term and long-term levels. Surely, potential endocrine disruption result in concerns at both levels? Regarding point 6, I am a bit concerned that we the level of knowledge regarding the normal histological profile of the developing tadpole along with the tissue specific gene expression profiles are generally lacking and therefore represent a major gap.

Another concern is the fact that we start the breeding by using high doses of hCG. Do we really understand the consequences of these doses for the mother (thyroid system) and the changes in aspects of maternal transfer, therefore impacting on the developing tadpole? This screening tool compare against a control, but maternal transfer may affect response sensitivities towards unknown compounds (false positives?).

5. **Impacts of the choice of test substances and methods chosen to demonstrate the performance of the assay**

A concern to me was that during this validation testing only limited potential mode-of-action modulation was tested. More attention should be give towards selecting controls representing different mode of actions, especially in a complex system like the thyroid. By including IOP in the Phase II series showed that at the developmental level unexpected results can be found. For this

assay we need causal relationships between morphological endpoints and different moed of actions. Moreover, I feel strongly that the link to possible use of the AMA in screening mixtures, and environmental samples at the Tier I level examples must be made. To what extend could the AMA be used to screen these complex samples?

The range of both agonistic and antagonistic representatives operating at different input sites (different modes of action) was rather limited and questions remain.

Repeatability and reproducibility of the results obtained with the assay, considering the variability inherent in the biological and chemical test methods

OECD Phase I study:- The repeatability of the AMA among three different laboratories showed some consistency. The outcome of the Phase I study corresponded to predicted results, especially in the higher concentration exposure groups. In this comparative study the fact that similar results were obtained in spite of variation in protocols used, show / confirm that the *Xenopus laevis* metamorphosis assay (XEMA) is a robust assay. Control compounds affected the thyroid histology as predicted. PTU exposure showed that chemicals affecting the iodine transport system will noticeably inhibit the TH output and thereby affect the functional aspects of the thyroid. Moreover, TH will result in increased thyroid activity. Compared to the Control tadpoles, significant inhibition and stimulation occurred. Differences in protocols used among the three participating laboratories clearly suggest that in spite of these differences, comparable results could be generated.

Multi-chemical study (USEPA):- The outcome of the known control chemicals (T4 and PTU) supported the results of the Phase 1 study. The results therefore confirm repeatability using the AMA protocol. In this study it was difficult to find the motivation (reason for inclusion) of most of the other chemicals (Discussion of Appendix H). The results of this study show that in many cases the understanding of the mode of action of lesser known chemical affecting the thyroid axis will need multiple endpoint studies. It is clear that the AMA represents a starting point only. In this study the importance of using the histopathology endpoints was underlined. One aspect that worried me is the chemical application (for example PCN). It seems that although theoretically sound it may be difficult for consulting laboratories to do exposure studies. Too little information is given on this aspect. I am still not convinced that body weight is a good indicator of thyroid effects. In this study histological observations were valuable and it shows that a combination of endpoints must be used. However, the studies present histopathology results as descriptions and it is difficult for the reader to visualize disruption. The question remains, how practical will it be for a reasonable inexperienced research team to evaluate histo-pathological endpoints? Another concern is that it was not clear whether the 21 day exposure starting with day 51 tadpoles were the better option (especially when evaluating histo-pathology). The compensatory hypothesis surfaced and more research is needed on this aspect. I am a bit worried that the limited number of endpoints, and the mode of action associated with these endpoints were not adequate to show thyroid disruption in some of the selected chemicals.

Inter-laboratory study (Phase II):- In this study all the knowledge and experience gained from the previous studies were used to standardize protocols. To solve some of the reproducibility issues more detailed descriptions of protocols were used. However, I think it is still not well-described and would lead to measurement errors and increased variation. A second concern that was highlighted was the variation that occurred in the development of control animals in one laboratory. The report attributes this variation to differences in feeding regimes. Added to this is the staging at of stage 51 tadpoles; could it be that inaccurate stage determination (stage 51) result in different developmental stages as early as day 7 of the exposure? Although the Perchlorate

control gave reasonable consistent results I was surprised by the inter-laboratory variation in results. I am not sure these variations were adequately addressed. One question that comes to mind is the aspect of observer error or reproducibility. Was the scoring of observers validated internally and between laboratories? The histological reading and scoring could be great source of variation. In general it seems that Perchlorate could be used as a standard control. How much regarding thyroid axis disruption can we read into general morphological endpoints like body size and weight? Just in control tanks we see so much variation in these growth parameters. The developmental endpoint in the Thyroid exposures corroborated previous studies and showed that this positive control worked well. However, the inconsistent results using the histological criteria were somewhat surprising. To what extend could this result be attributed to the fact that tadpoles were selected for histology on a random basis, therefore potentially including different NF stage tadpoles in the sample? Are we assuming that the histological picture is independent of developmental stage in exposed groups? Stage matched comparisons would have help answering this question. In the IOP control exposure the asynchronous development showed that staging problems may arise with certain chemicals. The question is would the gene expression studies (short-term study proposed by German group at som point) not helped to explain some of these results. I just get the feeling if endpoints respond strange or not at all in a limited array of endpoints, so much are lost. In this case the histology did not respond clearly either. It seems that if the mode of action is largely unknown then unpredicted results will make interpretations difficult.

The conclusions of the **Phase I and II** studies seems valid and underline certain concerns mentioned earlier. One aspect that increases the work load is the inclusion of a day 7 sampling. It seems that in the agonistic exposure (T4) growth parameters showed some sensitivity and helped interpretations when later compensatory effects came into play. However, whether the histological investigations at this stage made a valuable contribution was not clear. In general day 7 data seems to help the researchers to make an early assessment of how the exposure is going and it seems that the sampling at this time could be limited to save on labor.

OECD Phase III:- The stated goal of this exposures was to establish whether AMA could effectively indicate whether a compound needs further testing at Tier 2 level. The selection of compounds, for example 17β-Estradiol was not well-motivated. The statement that it is a potent endocrine disruptor is very general. To me endocrine disruption point to a mode of action or specific functionality and to include E2 only because it is a potent estrogenic EDC does not really make any sense. I presume the goal was to screen chemical with low predictive thyroid activity, but high activity in other areas of endocrine disruption? Was E2 included as a control since there is some indication that Benzophenone (BP2) is estrogenic? In the BP2 exposure study it was concerning that the two labs gave different results. It was attributed to differences in iodide concentration in the water. This underlines the value of standardizing all aspects of exposure when doing an inter-laboratory study. It was not clear why the difference in dilution water?

Other published studies:-From the literature it seems that results of known control chemical corroborate the results of the inter-laboratory studies, although in most cases histological studies were excluded from these.

Overall-comparison and Conclusions: -I suppose most data suggest that when using certain control chemicals (T4, PTU…) that the reproducibility of the AMA as a screening tool has been well demonstrated. This was especially true in the Phase I and II studies. Concerning was that not all aspects were always controlled for. Moreover, when conducting the inter-laboratory study using weak thyroid modulators, it seems that the consistency was lost.

The result of the inter-laboratory studies was the formulation of clear performance criteria. I agree it would reduce variability and ensure some form of assessment regarding performance of the metamorphosis assay. However, little attention was given to the source and time in captivity of the *Xenopus laevis* breeding pairs that a laboratory may use. Minimum median developmental stage of controls at the end of test may not be reached but the comparison between controls and experimental (unknowns) could still suggest further testing (Tier 2). The screening of the chemical is the main goal. Another question that should be asked: Is it necessary to include known agonist and antagonist controls? The implication is that the test laboratory always starts with three or four exposure groups. It seems that a laboratory could run these controls to determine capacity but that once this has been shown these could be exclude. The suggestion is that the performance criteria are applied after the 21 day trial. It seems from the studies conducted that one could include day 7 as some indicator? What about putting in a developmental check in the Control group at 14 days as well? To run the test to its completion and then assess performance seems unrealistic.

Overall utility of the assay as a screening tool to identify chemical that have the potential to interact with the endocrine system.
As pointed out in the objectives, the AMA as an *in vivo* screening tool represents a multi-endpoint model system. This assay integrates effects. Its greatest drawback is the time factor. Most organizations or researchers interested to screen compounds for more definitive testing are focusing on rapid tests, receptor binding assays or specific biochemical elements in certain pathways. From this perspective, the AMA is a long and labor intensive (expensive) bio-assay at the Tier I level. Indeed one may reason that we are paying high costs for an extensive complex bioassay with endpoints that are reasonable difficult to assess (especially the histological endpoints. However, the simplicity associated with the aquatic exposure of developing *Xenopus laevis* tadpoles offers unique opportunities to screen environmental chemicals. In contrast to mammals, tadpoles are assessable throughout their development and differential gene expression profiles exists throughout the developmental programme, making the selection of specific exposure windows more simple and controlled. Although the use of *in viv*o models for Tier I screening has been criticized it gives a more integrated response system. Therefore, I am convinced that the AMA has great advantages in identifying chemical that interact with the thyroid system. In combination with specific molecular end-points confident assessments will be made that will greatly aid the sorting of potential EDCs. Advancing to the screening of mixtures and environmental samples should be rather simple. The AMA lies at the interface of rapid, very sensitive and very specific in vitro assays, but with the advantage of an integrated in vivo response system, closer to the true picture of endocrine modulation. In addition, the AMA continues to contribute to the understanding of the role of thyroid hormone in vertebrate development, including mammals and humans. Since *X. laevis* has been a classical model system in embryology studies for decades, and the fact that several aspects of its endocrine physiology is well-understood together with recent advances made in the molecular field (creating specific tools to understand developmental stage-specific responses to TH and EDCs) the utility of the AMA assay is valuable and will allow for making links to more detailed studies regarding endocrine disruption.

In conclusion, I am of the opinion that the development and validation of the AMA using *X. laevis* as model has come a long way and should be implemented. However, it should be remembered that it is a qualitative screen. The refinements suggested should be incorporated and acknowledged that future refinements will continuously arrive to be incorporated. The AMA is a valuable and unique opportunity to use a rather simple *in vivo* system at the Tier I level.

Richard Wassersug Review Comments

REVIEW: The following text addresses each question of the official charge for the committee.

1. *Clarity of the stated purpose of the assay*.
The purpose of the Amphibian Metamorphosis Assay (AMA) is clearly stated in the EPA documents.

2. *Clarity, comprehensiveness and consistency of the data interpretation with the stated purpose of the assay*.
I have concerns about the comprehensiveness and consistency about the interpretation of the data from the various Phase 1, 2, and 3 trials. Almost all of my concerns center on the biology of the anuran larvae and the precision in which the assays were executed in the various labs. These concerns, as they arise in the "AMA Integrated Summary Report" of October 16, 2007, and are listed in order below.

Section 2.1—The first paragraph on the purpose of the assay states that "amphibian metamorphosis provides a well-studied thyroid-dependent process." In truth this has been studied in less than a dozen genera, out of the nearly 400 genera currently recognized in the Amphibia. I think it is important that the EPA documentation for the AMA include a introductory statement on the phylogenetic distance of the genus *Xenopus* from most other anurans, although that is touched upon in one of the background documents [notably ENV/JM/MONO(2004)17].

There is a tendency for molecular biologists, endocrinologists, and toxicologists to believe that what is true of *Xenopus* is true of *Rana*, and that what is true of both of those genera is true for all anurans. This presumption, for instance, is implicit in the introduction to Yun-Bo Shi's 2000 book Amphibian Metamorphosis and in the recent review by Fort et al. in Critical Reviews in Toxicology. In contrast, there are data indicating a variety of ways that *Xenopus* and *Rana* tadpoles differ.

One that may be particularly important to the AMA is how injured tails of these tadpoles respond to a retinoic acid challenge. Most anurans will start to differentiate hind legs and pelvic girdle at this caudal site of injury. *Xenopus*, however, does not show this response.

I would like comment on the appropriateness of *Xenopus laevis* as a model species for anuran metamorphosis assay. The various documents, particularly ENV/JM/MONO(2004)17, review the pros and cons of using *Xenopus*, particularly *X. laevis*, as the model species in the AMA. But they miss a few important points.

X. laevis has managed to establish itself as a feral exotic species on several continents outside of Africa, and can now be found in various disturbed and natural environments (e.g., in California, England, Chile) far beyond its normal range. As such the species appears to be exceptionally robust and tolerant of chemical stressors
(see http://www.columbia.edu/itc/cerc/danoff-burg/invasion_bio/inv_spp_summ/xenopus_laevis.htm). Thus, as acknowledged in the EPA and VOECD documents, a negative response from the AMA with *X. laevis* does not mean that a particular agent does not have a detrimental effect on other Anura. Several studies with the agents used in the Phase 1, 2, or 3 trials, which have yielded a positive response in *X. laevis*, have failed to do so at comparable doses in other species, or vice versa (e.g., see Ortiz-Santaliestra, 2007).

The value of the AMA is obvious when one is exploring compounds of a retinoid nature (cf. Gardiner et al., 2003; Fort et al., 2007). But what seems to be too often either hidden or forgotten is that how the

Xenopus response to retinoids is quite different than that of many other anurans. This is partially recognized in Degitz et al. (2000), but not in Degitz et al. (2003).

Tadpoles of many species will have a homeotic transformation of their tail tips into limbs, if the tail is injured and then challenged with retinoids. As Maden (1993) points out, this does not happen with *Xenopus* and I have personally confirmed that. I witnessed, however, that *Xenopus* growth was greatly retarded with retinoic acid. With high concentrations of retinoic acid, the tadpoles appear to starve to death (see also Degitz et al., 2003).

Xenopus and *Rana* have fundamentally different anatomy, functional morphology, and growth patterns for their tails and this may, in part, account for the different responses they have to the injury and to a retinoic acid challenge (Nishikawa and Wassersug, 1988). *Xenopus* tails continually add myotomes throughout the larval period, whereas *Rana* tails have deterministic growth (Wassersug 1997). As an aside, in a few studies where *Rana* have failed to show the homeotic transformation, I suspect that the stage of the tadpole and the dosage of the retinoic acid were not appropriate to elicit the response (in agreement with Degitz et al., 2000). This does not moot the utility of using the AMA for studying environmental retinoids. But it serves as a warning on how much one can safely infer from a negative response from *X. laevis* in the AMA.

This does not mean that *Xenopus* is not the best species for the AMA, but it does suggest that more emphasis should be given to encouraging researchers to be ready to explore further when a suspected endocrine disruptor yields a negative assay with *Xenopus*.

Although the concern just raised does not necessarily warrant changing the assay, but it does warrant changing the text. Thus for example, in the second to last paragraph in section 2.1, we are told that "the AMA focuses on anuran metamorphosis because it has been well-characterized." It would be more judicious to be a little more conservative and state that "the AMA focuses on *Xenopus* metamorphosis as a well-characterized example of metamorphosis in the Anura."

Section 3.3—This is the first place where we are introduced to hind limb length as a core endpoint of the AMA. Although this will seem like a trivial point, it may be worth specifying whether the left or right limb should be measured. I doubt that a lateralized difference in the length of the limbs occurs in *Xenopus*, but that is not impossible, considering that handedness in humans can affect aspects of limb size and development, and anurans do show handedness in hindlimb use (Robins et al., 1998).

Another point to consider is whether the AMA should request that both limbs be measured since increased fluctuating asymmetry is a well-established indicator of stress and disturbance during development in many vertebrates (http://www.animalbehavioronline.com/fa.html). Furthermore fluctuating asymmetry has been documented in the anuran appendicular skeleton (e.g., Vershinin et al., 2007; Söderman et al., 2007).

Section 3.4—The paragraph at the beginning of this section does not specify the size of the tanks in the various trials nor whether the flow-through system has the tanks in parallel or series. Granted, in the full description of the AMA both tank size and the flow-through path are specified, but it should be in the Summary as well.

Section 3.6—A key sentence in this section says that compounds which "are thyroid inactive will not likely undergo further testing to characterize thyroid activity." My concern here is that iodinase activity can occur even in frog embryos and can affect nervous system development (Dubois et al., 2006). Hence, anuran embryos can have a component of thyroid activity even before they actually have a thyroid gland. This suggests that thyroid function can be disrupted in an anuran even without a thyroid gland! This caveat

suggests that the FETAX assay should be considered and possibly run concurrently in certain cases where the AMA is also being used.

[As a small note, for some strange reason in this section, and in many tables the abbreviation "HHL" is used for hindlimb length. Elsewhere the more logical abbreviation "HLL" is used.]

Section 4.1—The rationale for using live brine shrimp as a food for *Xenopus* tadpoles in the US labs is not provided. *Xenopus* tadpoles are obligatory microphagous suspension feeders (Seale et al., 1982). It would be interesting to know whether the lab that fed its tadpoles live brine shrimp had evidence that the brine shrimp were effectively digested. If that diet improved tadpole growth, might it have been due to the addition of salt to the water along with the brine shrimp, rather than the shrimp themselves?

In the same section it is mentioned that "test vessel size and tank dimensions were not reported." Could the labs be contacted and asked for that missing information? To accept such important information as 'missing' is problematic. There is reason to believe that for a social schooling taxon like *X. laevis*, the number of tadpoles per volume can affect the growth rates, even when the food is abundant (see Katz et al., 1981). Thus without information on the density of the individuals (and not just the numbers per tank), it is not possible to fully interpret the different results among the different labs.

There is a problem with the anatomical terminology in the section where the thyroid gland histology is described. Here we are told that "transverse sections of the lower jaw" were made. If one is precise about the language, then none of the labs that did that would have found any thyroid issue. This is simply because the thyroid glands in *Xenopus* tadpoles lie within the brachial baskets, caudal to the "lower jaw," as shown in Fig. 1 in the "Guidance Document on Amphibian Histology Part 2." Obviously the various labs took not just the lower jaw, but the whole buccal floor and part of the pharyngeal (branchial) baskets; i.e., the floor of the mouth and the throat. This may seem like a petty point of language, but since there are few pictures in the literature about where the thyroid glands actually lie in *Xenopus* and other tadpoles, it would be helpful in the EPA documents were precise in terms of their terminology about the anatomical location of the gland.

On page 27, just below table 4-2, we come to the first mention of the use of static versus flow-through systems. Here we run into what I consider the first serious problem with the AMA methodology.

The statement on that page acknowledges that tadpole development under static conditions could be greater than tadpoles raised in a flow-through system, even when the same amount of food is provided. This is not surprising since *Xenopus* lives in still water in the wild and, as documented below, tadpoles are stressed when raised in a current. The response of *X. laevis* larvae to currents was examined more than a quarter of a century ago, but that literature is ignored in all of the AMA documents.

The closest the background literature comes to acknowledging the problem is on page 68 of the ENV/JM/MONO(2007)23 document. There it states that possible problems with the "established flow-through exposure system…[in the Japanese lab]…may explain some of the slight differences [in results] in the control animal performance." Those "slight differences," though, were the greatest in the inter-laboratory comparisons.

Section 5.1—Elsewhere the potential problem of currents for adults is recognized. So, for example, we find on page 52 under Section 5.111, paragraph 92, the statement: "Since *Xenopus* live naturally in static environments, care is required when using the flow-through systems so that the flow does not disturb the frogs." Since adults are negatively buoyant, benthic, and of relatively large mass, they can easily resist displacement in a gentle current without exerting energy. The tadpoles are not so lucky. Because of their

neutral or positive buoyancy, pelagic life-style, and small mass, they cannot avoid being displaced by a current without expending energy swimming upstream.

The stress to tadpoles raised in currents has recently been investigated in a stream-associated species *Rana boyii*. Dr. Sarah Kupferberg (Questa Engineering, Richmond CA, skupferberg@pacbell.net) has unpublished data that *R. boyii* tadpoles, which are far better designed for handling currents than *X. laevis*, exhaust in a matter of minutes in a current of just 5 cm/s.

Both in terms of the phase trials that were undertaken and the final AMA protocol itself, I strongly encourage the EPA to include document that raising *X. laevis* tadpoles in a current has an inconsequential effects on their growth and development. If the highest concern of the AMA methodology is to provide a continuous dose of the chemicals being assessed, then a rationale should be provided for why that is of higher priority than trying to raise the tadpoles in a slightly more natural and less stressful (i.e., in a non-flow-through) system. If a flow-through system is absolutely required, then much greater detail needs to be provided about the position of the inflow aperture(s) and whether it (they) induce a standing circulation in the tanks. In the current version of the AMA methodology, there is inadequate information on the permissible flow velocity in the tanks. Do the tadpoles line up with their nose pointing towards the inflow? If so, they are showing a positive rheotropic response and will be swimming harder (and expending more energy) than they would be in a static system. In addition, almost any current will cause major, non-random distribution of suspended food particles (see Walks, 2007). How will that affect growth rates and the variance in growth rates for the tadpoles in a single tank?

I do not wish to see this issue delay putting the AMA online as an approved EPA assay. But I do not feel that the AMA methodology can be considered in final form until there is some hard data showing that the inflow current is not affecting the tadpoles' behavior and growth. Since there is no detail provided on the currents generated by the flow-through system in the various Phase 1, 2, and 3 trials, this reviewer does not know whether the variance in the results from the different labs is not largely due to inadequate control of that particular variable.

On page 29, we learn that different labs anesthetized the animals different numbers of time. The final AMA protocol recognizes this as stressful for the tadpoles. None of the phase trials, though, explore this potential variable.

Lastly, no information is provided on the O_2 concentrations in the tanks. So, again, we don't know what importance differences in water chemistry may have had that could account for the different results between the different labs.

The next major problems all center on how one recognizes overt distress in a *Xenopus* tadpole.

Several of the tables that summarize the results from the various tests have a section titled "Overt Toxicity." There are three variables of 'toxicity' listed in those tables that are not strictly morphological markers. These are 'abnormal behavior,' 'lethargy,' and 'reduced food consumption.' To a non-behaviorist, it would seem that none of the labs witnessed any problems at any time in terms of any of these variables. However, since there is no discussion about what is normal behavior for a *Xenopus* tadpole I doubt that many (any) of the labs attempted to assess those variables…or were fully aware of what to look for in terms of behavioral disturbance.

Let's consider first 'abnormal behaviors.' *Xenopus* tadpoles are obligatory air breathers (e.g., Wassersug and Murphy, 1987; Pronych and Wassersug, 1994, plus older literature cited therein). They may come up to the surface in normoxic water only two or three times an hour, but, if they are stressed, they reduce their aerial respiratory rates. I have anecdotally noted (Wassersug, 1996) that simply tapping on the side of

Xenopus aquaria can reduce the tadpoles' aerial respiratory rates for up to an hour. Suppression of activity and reduced aerial respiration are well documented in the literature for stressed tadpoles, but never mentioned in the AMA documents.

Since labs that did all of the phase trials do not discuss the procedures they undertook to reduce the stress on the tadpoles, my guess is that all of the tadpoles were somewhat stressed. The problems then, are, "How much?" and "Was it the same amount of stress in all labs?"

Let's take a look at specific *X. laevis* behaviors. *Xenopus* tadpoles normally swim at an approximate 45º angle in the water column. However, if they are in a current they reduce their lung use and lung volume. They then have a shift in their center of buoyancy and swim more horizontally. None of the labs reported on the angle or orientation of the tadpoles. So we cannot tell whether their swimming was "normal," as it would be in standing water, or "abnormal" as it would be if they were swimming against a current and had reduced lung volumes.

When *Xenopus* tadpoles swim faster, they incorporate more of their tail in generating a propulsive wave. However, the frequency of the tail beat changes very little at low to moderate speeds (Hoff and Wassersug, 1986; Wassersug, 1989). What then was the wave pattern in the tails of these tadpoles? No data are provided.

To simply say that the tadpoles were swimming normally and "not lethargic," because the tails were constantly waving, presumes that the tail beat is under neuronal control. *Xenopus* tadpole tail tips, however, can continue to beat in tissue culture media for hours to days. So simply to witness that the animals swimming does not mean that they had normal behavior, i.e., that they were not "lethargic."

What other behaviors might have been examined and scored to document abnormal behavior, stress, or distress? The buccal pumping rate would be an obvious one. But there is no evidence that any of the labs measured this. This, in turn, directs our attention to the other measure of overt toxicity; i.e., "reduced food consumption." How did the labs measure the rate of "food consumption" to know if it was normal or reduced? *Xenopus* tadpoles reduce their buccal pumping rate when in a suspension with a high concentration of food particles (Seale and Wassersug, 1979; Seale et al., 1982). This is understood to be an adaptive response that helps the tadpoles avoid clogging their suspension feeding mechanism (Wassersug and Murphy, 1987). The Phase 1 and 2 laboratories, then, might have measured buccal pumping rates as an indirect proxy of feeding activity. However there is no evidence any lab collected such behavioral data.

A more direct measure of food consumption is a change in particle concentration in the water column around the tadpoles. This can be measured directly with a cell counting system, such as an automatic particle counter, or the old fashioned way using a grided slide under the microscope. But, once again, there is no evidence that any of the labs actually measured changes in particulate matter in the water, so it is not clear how they could have concluded that 'reduced food consumption' did not take place (other than indirectly from the final size of the tadpoles).

Whereas it is only a matter of history to criticize what was or wasn't done in the various labs, what really matters now is what is going to be considered normal tadpole behavior for the AMA. If the AMA is going to include measures of 'overt toxicity' that include behavior, then there must be rigorous and clear guidelines about what behaviors should be observed, how they should be quantified, and what is considered normal. In many ways this is the biggest weakness in the AMA documentation.

Before leaving this section, there is a sentence on page 34 that is unclear. That is where we are told that "hind limb length measurements were less straightforward due to a heterogenous effect in the Japanese laboratory." What is a "heterogenous effect?"

The last paragraph in section 4.3 concludes—despite all of the undiscussed and uninterpreted variation in results between the labs—that "the model system is relatively robust and not subject to variation as a function of the test protocols employed." I frankly do not see how such a strong statement can be made when there is variation in the test results between the labs in either gross morphology or thyroid histology that remains unexplained.

As we proceed through the document and the reviews of the various trials in the various labs, this same problem re-emerges. Thus we see on page 41 the statement that "the inter-study variability for wet weight of controls was somewhat greater." This raises a suspicion for me that the animals were subjected to different levels of stress in the different labs, but not enough information is provided to determine what those stressors were. As we work our way through the various chemical agents, we get hints of more variance possibly in behavior that is unexplained. On page 46, we are informed of sedative effects from phenobarbitol. But were those effects similar in all the labs?

The statement on page 49 of "a finding contrary to expected" would seem to have warranted some effort to figure out the source(s) of the variance. Yet the source or sources are not explored in these documents.

Section 5.2—One more hint that things were not normal (or at least consistent across labs) even in the controls, is the size range of the *Xenopus*. The average maximum size for *X. laevis* tadpoles in the wild is 80mm (Wager, 1965). The maximum size for tadpoles according to ENV/JM/MONO(2004)17 is 60 mm (page 52), but some of the lab results suggest that control animals are metamorphosing well below that size. It is quite likely then that the laboratory stock that have been used in the various laboratories around the world have been subjected to some substantive artificial selection, as well as the fact that the tadpoles may have been raised under non-optimal conditions.

I maintained a *Xenopus* colony for some 30 years. Over the years I found a tendency, when trying to maintain stock, to keep the first animals that metamorphose after a breeding and discard the extra tadpoles. In a few generations, this can lead to a bias for small individuals that metamorphose at a smaller size. I see nothing in the AMA that discusses how to maintain uniformity, if not 'wild type' in the breeding stock used in the assay. That issue needs to be addressed in the AMA methodology. If it is not addressed, then it belies the key statement in the Introduction to this section that "it is also imperative to refine husbandry methods and other test factors to ensure optimal and consistent performance of controls."

Only two paragraphs later, we are informed that "less than optimal control performance occurred in two experiments during the study." Without any effort to trace down the cause of that sub optimal performance, there is no guarantee that the AMA methodology can be consistently executed.

Section 5.4—I consider the inter-lab variation in tadpole size presented on pages 55 and 56 high. What guarantee do we have that the AMA in the future will perform any more consistenly? My concern repeats itself as we go from one test chemical to another. Thus, in section 5.4, we are told that there were "no signs of overt toxicity" but, as noted above, its not clear that the labs looked for behavioral indicators of stress or toxicity. Considering the fact that T4 is a thyroid hormone, one would hope that the assay could run without the level of variability reported for T4 on pages 61 and 62.

When we learn that laboratory 5 had mortality "due to handling errors" warning lights go off in my head. What were the errors? Were all the animals abused, but only a few of them dying? Were those "handling errors" isolated and specifically involving only the individuals that actually died? Or were all the tadpoles exposed to those "handling errors" and some of them were hardy enough to survive?

Given the variation reported between the three labs, the last sentence on page 67 is bothersome. We are told that "the strong developmental response was deemed to be sufficient to conclude that the assays successfully detected T4." This seems to me a trivial statement, since we have known for decades that T4 affects the development of *Xenopus* tadpoles. What is so problematic is the begining of that sentence where we learn that "thyroid histopathology as inconsistent between the three labs." Thus, for certain agents in certain labs, histopathology is a powerful aspect of the AMA's ability to discern endocrine disruption. In other labs, histopathology yields inconsistent results. Without chasing down the source of such inconsistencies, we cannot have full faith that the AMA protocol can produce consistent results between different labs.

The problem keeps returning. So, on the bottom of page 70, we learn of an additional difference in results from the various labs for which "the reason…has not been determined." By now one has the impression that many months, in many labs, were spent to show the obvious; i.e., that compounds like T4, which are thyroid promoters, accelerate development whereas compounds that have long been known to inhibit metamorphosis, do so in more than one lab. Yes, the AMA works! But not ideally, and not consistently. So I'm left wondering why more effort was not put into trying to identify and resolve the variation reported in the results from the different labs.

The various sections all seem to end with some statement that the assay worked. Thus we are told that the strange development observed in the tadpoles in the iopanoic acid (IOP) studies (with "asynchronous development") simply because it gave a response "can be considered a 'positive' result." Yes, positive. But otherwise uninterpretable.

The last paragraph on page 76 states that the iodine content in the culture water "must be considered." It isn't clear that this was addressed in the earlier phase trials and may be more important than is appreciated in the current AMA methodology.

Section 5 ends with a statement that metamorphosis in *Xenopus laevis* could be used as a "testing tool for thyroid system disruption." While this important concluding statement is in italics, this was clearly known twenty years ago.

Section 6.2—Here is an aside on biology, and not on the assay *per se*. I found it intriguing that estrogen increased the size of the *Xenopus* tadpoles. Adult female *Xenopus* are larger than males. Over the years, I have been occasionally asked if there is some way to tell male from female tadpoles. It would be fun now to go back to the lab and find out whether, all else being equal, female *Xenopus* tadpoles are larger than males at or before metamorphosis. Hayes et al.'s (1993) failure to find any estrogenic effect on *Bufo* larval growth and metamorphosis doesn't moot the question. It is my impression that species, whose size at metamorphosis is closest to their size at first reproduction, are more likely to show differentiation of their gonads at metamorphosis than species that metamorphose at a size well below their reproductive size. Sex difference in size at metamorphosis may thus be most likely to be found in the former rather than the latter group.

Section 7—This section acknowledges that the scientific literature was reviewed up to 2003. It is not clear why the literature wasn't updated for the last three or four years. The literature, though, is updated in Fort et al. (2007).

In section 7.2, we are introduced to *Silurana* (*Xenopus*) *tropicalis* as alternative model species. We are told that it could be used in place of *Xenopus laevis* "with minimal modifications," but those modifications are never specified.

Section 8.1—This section proclaims "The reproducibility of the AMA, for screening purposes, has been well-demonstrated using several representative thyroid-active chemicals across geographically diverse laboratories." However, if the variation between the labs cannot be explained, then one cannot feel as confident about this proclamation as the author of the review.

Section 8.3—Here the strengths and limitations of the assay are listed. I agree with the combination of morphological and histological endpoints, but they are only considered acceptable within the context of the animals having normal behavior. Without defining 'normal behavior,' and without any clear guidance on how to quantify that, it is not clear how sensitive, reproducible, and reliable the AMA will be.

3. *Biological and toxicological relevance of the assay as related to its stated purpose.*
The AMA with *Xenopus* is toxicologically relevant in that this is the most common amphibian used in toxicological research around the world. Its biological relevance, however, is slightly less relevant in some situations. *Xenopus* is not native to any continent outside Africa, and its morphology, ecology and behavior both as a larva and adult, are quite unlike those of other amphibian genera in North America, Asia, Europe, or Australia. The authors of the EPA documents suggest that the agency is aware of situations where data collected with *Xenopus* via the AMA, may not be relevant for other species and mentions the potential need to verify the results from the AMA with other anuran taxa.

4. *Clarity and conciseness of the test method in describing the methodology of the assay such that a laboratory can a) comprehend the objective, b) conduct the assay, c) observe and measure prescribed endpoints, d) compile and prepare data for statistical analyses, and e) report results.*
My greatest concerns about the AMA center on the document "Draft Method for the AMA." Various laboratories should be able to follow the methodology of this essential document and achieve identical results. There is simply not enough detail in this methodology to be confident that the assays can be executed with adequate amounts of reproducibility.

The following is a list of my major concerns.

Breeding stock—No guidance is provided on whether one should be concerned about inbreeding in laboratory stock. As noted in Item 2, the labs in general seem to be reporting tadpoles in control tanks metamorphosing below the maximum size in nature. As I've noted above, it is easy to artificially select for larvae that metamorphose at a small size. But how would that affect the results of the AMA? One guess is that it would reduce the sensitivity of the assay. If a presumed endocrine disruptor reduces the size of tadpoles, and the tadpoles used in that assay have already been selected to be dwarfs, then it's going to be more difficult for the AMA to pick up a significant reduction in size.

I do not recommend that the EPA delay putting the AMA into operation. But ways to either deal with or avoid using inbred lines need to be addressed. Whatever their guidelines are, they have to be tight enough that they yield standardized breeding stocks across various labs.

Exposure system—Another major concern I have is with the mechanics of the flow-through dilutor system. I understand that the tanks will be in parallel, not in series, which, of course is essential. But much more information is needed to make sure that all the labs produce comparable circulation in their tanks by: 1) having identical placements of the inflow and outflow apertures, 2) apertures of identical size, 3) yielding identical flow rates and circulation in the tanks.

In Item 2 above, I emphasize that *Xenopus* tadpoles live in non-flowing water and that putting them in a current is stressful. The background literature in support of this claim goes back at least a decade, some of it to the early 1980s. Nowhere in these documents do I see those concerns mentioned or discussed.

Minimally the AMA should include ways to minimize the current velocity in the tanks, such that there will not be a major, standing circulation**.**

Please consider the following: *Xenopus* tadpoles in a current reduce their aerial respiration rate. They do this by lowering the volume of air in their lungs. This makes them more negatively buoyant so they can stay closer to the bottom, where the flow rate is lowest. This, however, lowers their stamina and can increase their lactic acid concentration (see Wassersug and Feder, 1983; Feder and Wassersug, 1984). If the lactic acid is elevated, then the animals are stressed. Stress increases corticotropin-releasing factor (CRF) which has been shown to activate both adrenal (interrenal) and thyroid hormone secretions (see Denver, 1996, plus other papers cited therem as well as reviewed in Wells, 2007, p. 608 and Fort et al., 2007).

What is remarkable is that the EPA documents fully acknowledge the problem of stress from an endocrinological perspective, yet completely ignore it from an ecological and behavioral perspective. Thus, in the ENV/JM/MONO(2004)17 document (which is in general an excellent document) we are told explicitly on page 27 that CRF, not TRH (=thyrotropin releasing hormone not "thyroid receptor element" as claimed in Table 1-1, page 20 of the same document) "is the primary hypothalamic releasing hormone responsible ultimately for the induction of metamorphosis." On the next page we learn that many tissues in tadpoles are responsive to the impact of corticoids on thyroid hormone action. The section ends (paragraph 29) with the statement that "Overall, physiological synthesis and secretion of corticoids play an important role in anuran metamorphosis." In layman's terms, these quotes recognize that the endocrinological pathways that respond to environmental stress interact with the endocrinological pathway that control metamorphosis. Yet the AMA documentation says nothing about how to limit, or even recognize and regulate non-chemical environmental stressors on tadpoles.

Since the EPA is committed to a flow-through system, in order to stabilize the delivery of the test compounds, far more effort needs to be spent on how to do this in a way that minimizes—or at least standardizes—the stress that currents, for example, place on *Xenopus* tadpoles.

Removing the jelly from the eggs—An optional step in the production of tadpoles for the AMA is to use L-cysteine to remove the jelly. It is not clear why this should be done, optionally or otherwise. From a historical perspective, one can understand why many labs do this. It is, for example, part of the FETAX, which is an assay for developmental disruptors of embryogenesis. Since the concern in that assay is to get the test agent to the embryo in a consistent fashion, it makes sense to remove the jelly, which may or may not be uniform on different eggs and may inhibit transfer of the test chemicals to the embryos themselves. Removing the jelly is also a step in all trangenic work with *Xenopus* eggs. However, in light of the concerns that iodine in the water may be an important variable that needs to be controlled, I feel that the L-cysteine step should not be optional.

A case can be made for removing the jelly to make sure that iodine and other growth promoting elements in the water (most notably O_2) are not blocked from getting to the embryo. Notably, this has relevance to the 'thyroid axis' even in the early embryo. Dubois et al. (2006) point out that thyroid hormone is assumed to be absent in embryos before they develop a differentiated thyroid gland. However, they show that elements of thyroid hormone signaling pathways are present during early development of *Xenopus*. They find, for example, functional deiodinase activity and even T4 at significant levels during early embryogenesis, this pre-thyroid gland hormonal activity is substantive in neurogenic areas.

An implication of the Dubois et al. study is that thyroid hormonal function can affect tadpole development long before the tadpoles reach NF stage 51. Without more knowledge about how the jelly affects this embryo biochemistry, a case can be made for removing it from all eggs to strive for better consistency.

[Minimally, those who run the AMA need to have control of iodine concentration in the water right from the time that they start breeding the adults, and not just during the execution of the AMA.]

There is, however, an alternative way of looking at this. If we are concerned about whether a certain agent is an endocrine disruptor in the natural environment, we should remember that frogs' eggs all have gelatinous coats in the wild, and this material may have a protective function for the embryos. If the results from the AMA are to be most meaningful for other species in the wild, a case could be made for leaving the jelly on, to help make the *Xenopus* eggs more comparable to those of other species in the wild.

Either way—with or without jelly—the EPA should arrive at a consistent and non-optional policy about how the eggs for the AMA should be raised.

Larval care and selection—The AMA similarly must come up with clearer guidelines on how to standardize, if not minimize, the daily disturbance to the tadpoles. In the Methods document there is only a single sentence on cleaning the tanks. There we are told that the tanks "shall be siphoned clean daily." There are no guidelines on how to do this in a standardized fashion that minimizes the stress on the tadpoles.

As mentioned above, tapping on the side of an aquarium can cause *Xenopus* tadpoles to reduce their aerial respiration rate, even when their swimming and other behaviors appear perfectly normal. Siphoning the bottom of a tadpole tank must surely be a comparable or more extreme stressor.

It is well known for tadpoles of other species that they retreat to the shallows and stay near the bottom when they sense a threat. Clearly, intensively siphoning the tank would be a stressful mechanical disturbance for any tadpole. Rot-Nikcevic et al. (2005) found that mechanical disturbance can indeed reduce the growth rate of *Xenopus* tadpoles. Although their data were not statistically significant at the $P < 0.05$ level, their mechanically disturbed *Xenopus* tadpoles were on average 10% smaller than undisturbed tadpoles.

Older data in Wassersug and Murphy (1987) show that aerial respiration facilitates growth in *Xenopus* larvae. Denying *Xenopus* access to air by stressing them so they avoid the air-water interface is likely to retard metamorphosis (Pronych and Wassersug, 1994). Feder and Wassersug (1984) show that 16.6% of the total O_2 consumption for *Xenopus* larvae in normoxic water comes from aerial respiration. This can increase to 100% in hypoxic water. All of these data suggest that mechanical disturbance is likely to negatively impact on *Xenopus* larvae in the AMA. This mechanical disturbance can be from cleaning activity, noise from pumps, human activity around the tanks, bubble stones or other aerating machinery, etc. In order for the AMA to yield consistent results between labs, the protocol must include rigorous standards for controlling, if not eliminating, these sources of stress to the tadpoles**.**

Establishing the highest test concentration—There is a subtle contradiction in the example given under the subheading of "test concentration range." There we are told that the minimal range "shall be at least one order of magnitude" but that is immediately followed by an example where the range runs from 0.11 to 1.0, which is slightly less than one full order or magnitude.

Daily observations of test animals—We are told this is necessary, but there are no directions about what one should be observing. Yet again, it seems imperative that the AMA define more rigorously what constitutes normal behavior for *Xenopus* tadpoles.

Hindlimb length—Should the same side of the tadpole be measured in all the labs? Should labs measure both sides so they can collect data on fluctuating asymmetry?

Body length and wet weight—More direction is necessary to standardize how one should remove adherent water from the body of tadpoles before their weight is determined. The document recognizes that "weight determinations can cause stressful conditions for tadpoles and may cause skin damage." This would mandate standardization in this step. Over the years I've watched students very gently pick up tadpoles with a dipnet and do virtually nothing to remove surface of water for fear of injuring the larvae. I've also seen tadpoles get shaken down vigorously and patted dry as if they were vegetables being prepared for a salad. The EPA needs to provide greater direction about how the tadpoles should be freed from surface fluid in order to increase the chances of comparable weight measurements between labs.

Additional observations—The text here makes it clear that the EPA expects behavior to be monitored, but it gives no guidance on how to do this. Taking each one of their examples, one can see problems.

They start off by mentioning "uncoordinated swimming." *Xenopus* is a social species. Is "uncoordinated swimming" then measured by the geometry of the school (e.g., orientation of one tadpole to another? distance between tadpoles? etc.). The distance between tadpoles varies depending on their size, density, and illumination (Katz et al., 1981). But chemical agents can also affect the interactive distance; i.e., the 'coordinated' nature of their swimming within a school (Lum et al., 1982). Should this be measured to determine if their swimming is coordinated?

The next variable mentioned is 'hyperventilation." Ventilation for *Xenopus* tadpoles has both an aerial and an aquatic component. Under normoxic conditions, tadpoles come to the surface to take air about twice an hour. If they were to come up three or four times an hour, that would be a 50 and 100% increase in their aerial respiratory rate and could be considered "hyperventilation."

One may suppose that the authors of the AMA protocol were not thinking about aerial respiration at all, but only aquatic ventilation. There is, however, still a problem. The primary determinant of buccal pumping rate (i.e., aquatic ventilation) is not O_2 concentration, but the density of particulate matter in the water (see Feder et al., 1984; Seale et al., 1982). Thus a "hyperventilating tadpole" may be experiencing hunger rather than respiratory distress. Without standardizing exactly when food is delivered to the tanks, how uniformly it is dispersed in the water, and how rigorously ventilation is measured, there will be no way for any lab to determine whether the tadpoles are indeed hyperventilating.

Next on the list is "atypical quiescence." I have no idea what that means or how it is supposed to be measured.

The last variable is "non-feeding," but again there is no indication of how that is supposed to be measured. *Xenopus* tadpoles can regularly feed on suspended particles that are too small to be seen with the naked eye; they are continuous, obligatory, suspension feeders. If they were not trapping particles in mucous, the particles would be going into the mouth through their gill slits and out again. They would then be "non-feeding." But how would any lab determine that?

Possibly the author(s) of the AMA protocol expect those using the AMA to be measuring buccal pumping rate. That is the only variable which can be easily measured that is an indirect behavioral proxy for whether a tadpole is feeding or not. But there are no guidelines provided about how and when to do this.

O_2 concentration—The AMA sets a range for O_2 concentration which should be no less than 40% of air saturation. It does not specify how the water should be aerated in order to maintain that concentration. That needs to be standardized in order to reduce disturbance to the tadpoles.

Water temperature—The water temperature is supposed to be maintained at 22 +/– 1 ºC. This is slightly above preferred room temperature for North America, which is usually 21ºC. With air temperature of

precisely 21ºC, evaporative cooling would lower the water temperature to slightly below the 22 +/– 1 ºC range. That would then require some way of heating the water to bring it up to 22+/- 1 ºC. How is that temperature supposed to be maintained?

There are various options for maintaining the tank temperature above room temperature. They range from individual heaters in the tanks to heating the water in the up-stream reservoir for the flow-through system.

I did not see documentation on how different the growth would be for the *Xenopus* tadpoles, if they were raised, say, at 21.1 versus 22.9 ºC, even though both would be in the acceptable range of 22 +/– 1 ºC. It is not clear how the range of +/- 1 ºC was established. One suspects that it was simply convenient and not based on firm data to show that there were no differences in the growth and metamorphosis of *Xenopus* at 21.1 ºC versus 22.9 ºC. In a flow-through system, it can be difficult to maintain thermal constancy within a tank. More guidance should be provided about how to stabilize the temperature in the tanks.

5. *Strengths and/or limitations of the assay.*
The greatest strength of the assay come from the amount of work that the EPA, its partners and its contractors have put into developing the assay over the last decade. They have made major progress in developing a reliable amphibian metamorphosis assay. Given the concerns about endocrine disruptors in the environment, this effort was appropriate. There are, however, some holes in the protocol about how to perform the assay. As stated extensively above, important variables in the execution of the assay are missing from the documents provided. The biological relevance has to be qualified given how different *Xenopus* is than all of the non-pipid anurans in the world (see #3 above).

6. *Impacts of the choice of test substances and methods chosen to demonstrate the performance of the assay.*
The tests subjects used to demonstrate the performance of the assay were appropriate as were most of the methods used. However there are still some methodological problems, which are discussed extensively above.

7. *Repeatability and reproducibility of the results obtained with the assay, considering the variability inherent in the biological and chemical test methods.*
One of my greatest concerns in the AMA documentation is the high variance in reproducibility of the results obtained from the various labs during the various test phases. I am disquieted by the little attention given to the variance between the labs, when their protocols were (supposedly) identical.

Most of the chemicals used in these studies were well known inhibitors or accelerators of metamorphosis. The fact that inhibition and acceleration were seen in the test results is, of course, exactly what one expected. I did not expect, however, the variance in the reports between the different labs. It is bothersome that more effort was not made to explain the inter-laboratory variance.

8. *Please comment on the overall utility of the assay as a screening tool, to be used by the EPA, to identify chemicals that have the potential to interact with the endocrine system to warrant further testing.*
Despite all the concerns stated above, I feel that the EPA should accept the AMA—*with expansion of its protocol documentation*—as a screening tool for chemicals that may have the potential to interact with the vertebrate endocrine system. I encourage the EPA to proceed with putting this assay on-line, while they concurrently address the many concerns raised in Items 2 and 4 above.

REFERENCES

Degitz, S.J., P.A. Kosian, E.A. Makynen, K.M. Jensen and G.T. Ankley 2000 Stage- and species-specific developmental toxicity of all-trans retinoic acid in four native North American ranids and *Xenopus laevis*. *Toxicol. Sci*., 57:264-274.

Degitz, S.J., E.J. Durham, J.E. Tietge, P.A. Kosian, G.W. Holcombe and G.T. Ankley 2003 Developmental toxicity of methoprene and several degradation products in *Xenopus laevis*. *Aquat. Toxicol.*, 64:97-105.

Denver, R.J. 1996 Neuroendocrine control of amphibian metamorphosis. In: "*Metamorphosis: Post-Embryonic Reprogramming of Gene Expression in Amphibian and Insect Cells*." J. R. Tata, L. I. Gilbert and E. Frieden (eds.) Academic Press, Orlando, pp. 433-464.

Dubois, G.M., A. Sebillot, G.G.J.M. Kuiper, C.H.J. Verhoelst, V.M. Darras, T.J. Visser and B.A. Demeneix 2006 Deiodinase activity is present in *Xenopus laevis* during early embryogenesis. *Endocrinology*, 147:4941-4949.

Feder, M.E. and R.J. Wassersug 1984 Aerial versus aquatic oxygen consumption in larvae of the clawed frog, *Xenopus laevis*. *J. Exp. Biol.*, 108:231-245.

Feder, M.E., D.B. Seale, M.E. Boraas, R.J. Wassersug and A.G. Gibbs 1984 Functional conflicts between feeding and gas exchange in suspension-feeding tadpoles, *Xenopus laevis*. *J. Exp. Biol.*, 110:91-98.

Fort, D.J., S. Degitz, J. Tietge, L.W. Touart 2007 The hypothalamic-pituitary-thyroid (HPT) axis in frogs and its role in frog development and reproduction. *Crit. Rev. Toxicol.*, 37:117-161.

Gardiner, D., A. Ndayibagira, F. Grun and B. Blumberg 2003 Deformed frogs and environmental retinoids. *Pure Appl. Chem*., 75:2263-2273.

Hayes, T.B., R. Chan and P. Licht 1993 Interaction of temperature and steroids on larval growth, development and metamorphosis in a toad (*Bufo boreas*). *J. Exp. Zool*. 266:206-215.

Hayes, T.B. 1997 Steroid as potential modulators of thyroid hormone activity in anuran metamorphosis. *Am. Zool*., 37:482-490.

Hoff, K. and R.J. Wassersug 1986 The kinematics of swimming in larvae of the clawed frog, *Xenopus laevis*. *J. Exp. Biol*., 122:1-12.

Katz, L., M. Potel and R.J. Wassersug 1981 Structure and mechanisms of schooling in larvae of the clawed frog, *Xenopus laevis*. *Anim. Behav.*, 29:20-33.

Lum, A., R.J. Wassersug, M. Potel and S. Lerner 1982 Schooling behavior of tadpoles: A potential indicator of ototoxicity. *Pharmac. Biochem. Behav*., 17:363-366.

Maden, M. 1993 The homeotic transformation of tails into limbs in *Rana temporaria* by retinoids. *Dev. Biol.*, 159:379-391.

Nishikawa, K. and R.J. Wassersug 1988 Morphology of the caudal spinal cord in *Rana* (Ranidae) and *Xenopus* (Pipidae) tadpoles. *J. Comp. Neurol*., 269:193-202.

Ortiz-Santaliestra, M.E. and D.W. Sparling 2007 Alteration of larval development and metamorphosis by nitrate and perchlorate in Southern Leopard Frogs (*Rana sphenocephala*). *Arch. Environ. Con. Tox.*, 53:639-646.

Pronych, S. and R.J. Wassersug 1994 Lung use and development in *Xenopus laevis* tadpoles. *Can. J. Zool*., 72:738-743.

Robins A., G. Lippolis, A. Bisazza, G. Vallortigara and L.J. Rogers 1998 Lateralized agonistic responses and hindlimb use in toads. *Anim. Behav*., 56:875-881.

Rot-Nikcevic, I., R.J. Denver and R.J. Wassersug 2005 The influence of visual and tactile stimulation on growth and metamorphosis in anuran larvae. *Funct. Ecol*., 19:1008-1016.

Seale, D.B. and R.J. Wassersug 1979 Suspension feeding dynamics of anuran larvae related to their functional morphology. *Oecologia*, 39:259-272.

Seale, D.B., K. Hoff and R.J. Wassersug 1982 *Xenopus laevis* larvae (Amphibia, Anura) as model suspension feeders. *Hydrobiologia*, 87:161-169.

Söderman, F., S. Dongen, S. Pakkasmaa and J. Merilä 2007 Environmental stress increases skeletal fluctuating asymmetry in the moor frog *Rana arvalis*. *Oecologia*, 151:593-604.

Vershinin, V.L., E.A. Gileva and N.V. Glotov 2007 Fluctuating asymmetry of measurable parameters in *Rana arvalis*: Methodology. *Russ. J. Ecol+.,* 38:72-74.

Wager, V.A. 1965 *The Frogs of South Africa*, Purnell & Sons, Johannesburg.

Walks, D.J. 2007 Persistence of plankton in flowing water. *Can. J. Fish. Aquat. Sci*., 64:1693-1702.

Wassersug, R.J. and M.E. Feder 1983 The effects of oxygen concentration, body size and respiratory behaviors on the stamina of obligate aquatic (*Bufo americanus*) and facultative air-breathing (*Xenopus laevis* and *Rana berlandieri*) anuran larvae. *J. Exp. Biol*., 105:173-190.

Wassersug, R.J. 1996 The biology of *Xenopus* tadpoles. In: R.C. Tinsley and H.R. Kobel (eds.). *The Biology of Xenopus*, Clarendon Press, Oxford, pp. 195-211.

Wassersug, R.J. 1989 Locomotion in amphibian larvae (or "Why aren't tadpoles built like fishes?"). *Amer. Zool*., 29:65-84.

Wassersug, R.J. and A.M. Murphy 1987 Aerial respiration facilitates growth in suspension-feeding anuran larvae (*Xenopus laevis*). *Exp. Biol*., 46:141-147.

Wassersug, R.J. 1997 Where the tadpole meets the world-Observations and speculations on biomechanical and biochemical factors influencing metamorphosis in anurans. *Amer. Zool*., 37:124-136.

Wells, K.D. 2007 *The Ecology and Behavior of Amphibians*, University of Chicago Press, Chicago, pp. 608.

Shi, Y. 2000 *Amphibian Metamorphosis*, John Wiley & Sons, Inc., Toronto.

Appendix A

CHARGE TO PEER REVIEWERS

INDEPENDENT PEER REVIEW OF THE AMPHIBIAN METAMORPHOSIS ASSAY AS A POTENTIAL SCREEN IN THE ENDOCRINE DISRUPTOR SCREENING PROGRAM (EDSP) TIER-1 BATTERY

October 16, 2007

Background:

According to Section 408(p) of the EPA's Federal Food Drug and Cosmetic Act, the purpose of the EDSP is 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 Act, the EPA formed the Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC), a panel of scientists and stakeholders that was charged by the EPA to provide recommendations on how to implement the EDSP. Upon recommendations from the EDSTAC, the EPA expanded the EDSP using the Administrator's discretionary authority to include the androgen and thyroid hormone systems, as well as wildlife.

One of the test systems recommended by the EDSTAC was the Amphibian Metamorphosis Assay (AMA). The AMA consists of multiple endpoints; principally, developmental stage, hind limb length, body length (whole body and snout-vent), histology of the thyroid glands, mortality and morbidity. It is intended to empirically identify substances which may interfere with the normal function of the hypothalamic-pituitary-thyroid (HPT) axis. It represents a generalized vertebrate model to the extent that it is based on the conserved structure and functions of thyroid systems. It is an important assay in the EDSP screening battery because amphibian metamorphosis provides a well-studied, thyroid-dependent process which responds to substances active within the HPT axis, and it is the only assay in the battery that assesses thyroid activity in an animal undergoing morphological change. The AMA is not intended to quantify or confirm endocrine disruption, or to provide a quantitative assessment of risk, but only provide suggestive evidence that thyroid regulated processes may be sufficiently perturbed to warrant more definitive testing. A weight-of-evidence approach among the multiple endpoints within the bioassay, combined with biological plausibility, is expected to help distinguish endocrine-related effects from spurious effects and to determine whether a chemical substance has a positive endocrine effect on the HPT axis.

Although peer review of the AMA will be done on an individual basis (i.e., its strengths and limitations evaluated as a stand alone assay), this assay, along with a number of other *in vitro* and *in vivo* assays, will likely constitute a battery of complementary screening assays. A weight-of–evidence approach will also be used *among* assays within the Tier-1 battery to determine whether a chemical substance has the potential to interact with the endocrine system and whether Tier-2 testing is necessary. Peer review of the EPA's recommendations for the Tier-1 battery will be performed at a later date by the FIFRA Scientific Advisory Panel (SAP). The battery peer review will focus, in part, on the issue of coverage of known modes of endocrine activity, and how well individual assays work in concert with one another within the Tier-1 battery. Hence, it is important to peer review each individual assay.

Each peer reviewer is asked to review the Integrated Summary Report and draft test method, with accompanying support materials, and comment on the results of the validation process of the AMA,

with special attention given to the inter-laboratory validation exercise. Review and comment shall be directed to each of the following:

1. Clarity of the stated purpose of the assay.

2. Clarity, comprehensiveness and consistency of the data interpretation with the stated purpose of the assay.

3. Biological and toxicological relevance of the assay as related to its stated purpose.

4. Clarity and conciseness of the test method in describing the methodology of the assay such that a laboratory can:

    a. comprehend the objective,
    b. conduct the assay,
    c. observe and measure prescribed endpoints,
    d. compile and prepare data for statistical analyses, and
    e. report results.

    If warranted, please also make suggestions or recommendations for test method improvement.

5. Strengths and/or limitations of the assay.

6. Impacts of the choice of test substances and methods chosen to demonstrate the performance of the assay.

7. Repeatability and reproducibility of the results obtained with the assay, considering the variability inherent in the biological and chemical test methods.

8. Please comment on the overall utility of the assay as a screening tool, to be used by the EPA, to identify chemicals that have the potential to interact with the endocrine system sufficiently to warrant further testing.

Appendix B

INTEGRATED SUMMARY REPORT

Integrated Summary Report for Validation of a Test Method that Assesses the Potential of Chemicals to Interfere with HPT Axis Structure and Function, as a Potential Screen in the Endocrine disruptor Screening Program Tier 1 Battery (PDF) (95pp, 774 K)

Appendix C

SUPPORTING MATERIALS

Attachment A: Test Method for the Amphibian Metamorphosis Assay (ZIP) (3 files, 4.8M)

Test Method for the Amphibian Metamorphosis Assay (PDF) (26pp, 224K)

Appendix 1 for the Test Method of the Amphibian Metamorphosis Assay - Guidance Document on Amphibian Thyroid Histology Part 1: Technical Guidance for Morphologic Sampling and Histological Preparation (PDF) (16pp, 215K)

Appendix 2 for the Test Method of the Amphibian Metamorphosis Assay - Guidance Document on Amphibian Thyroid Histology Part 2: Approach to Reading Studies, Diagnostic Criteria, Severity Grading, and Atlas (PDF) (47pp, 4.24M)

Attachment B: Final Report of the Validation of the Amphibian Metamorphosis Assay for the Detection of Thyroid Active Substances: Phase 1, Optimisation of the Test Protocol

Final Report for Phase 1 of the Validation of the Amphibian Metamorphosis Assay (PDF) (90pp, 717K)

Attachment C: Final Report of the Validation of the Amphibian Metamorphosis Assay: Phase 2 - Multi-Chemical Interlaboratory Study

Final Report for Phase 2 of the Validation of the Amphibian Metamorphosis Assay (PDF) (96pp, 790K)

Attachment D: Draft Report of the Phase 3 of the Validation of the Amphibian Metamorphosis Assay

Draft Report of the Phase 3 of the Validation of the Amphibian Metamorphosis Assay (PDF) (73pp, 320 K)

Attachment E: Detailed Review Paper for the Amphibian Metamorphosis Assay

OECD Detailed Review Paper for the Amphibian Metamorphosis Assay (PDF) (106pp, 522K)

Attachment F: Endocrine Disruptor Screening Program Validation Paper

Paper Describing EDSP's Approach to Validation for the Tier 1 Screening Battery (PDF) (21pp, 131K)

Attachment G: Power Analysis for Determining Study Design for the Amphibian Metamorphosis Assay

Power Analysis Within the Proposal for Phase 1 of Validation of the Amphibian Metamorphosis Assay (PDF) (66pp, 1.72M)

Attachment H : Final Report (Battelle) for the Multi-Chemical Study (ZIP) (2 files, 32.5M)

Final report for the Multi-Chemical Study Performed by a Contractor for the Amphibian Metamorphosis Assay Validation (PDF) (54pp, 355K)

Appendices to the Final Report (Battelle) for the Multi-Chemical Study (PDF) (903pp, 34.7M)