Peer Review Results for the 15-Day Intact Adult Male Rat Assay

Prepared for:

U.S. Environmental Protection Agency
Exposure Assessment Coordination and Policy Division
Office of Science Coordination and Policy
1200 Pennsylvania Avenue, N.W.
Washington, DC 20460

Prepared by:

Eastern Research Group, Inc.
14555 Avion Parkway
Suite 200
Chantilly, VA 20151-1102

04 October 2007

EPA Contract No. EP-W-05-014
Work Assignment 3-13

TABLE OF CONTENTS

LIST OF TABLES

Page

1.0 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 15-day intact adult male rat assay. The intact adult male assay consists of multiple endpoints; principally, terminal weights of primary and secondary sex organs and thyroid gland, histology of the testes, epididymides and thyroid, and serum concentrations of reproductive steroids, gonadotropins and thyroid hormones.

According to numerous reports published in peer-reviewed scientific journals, the intact adult male rat assay has the capacity to detect estrogen receptor agonists/antagonists, androgen receptor agonists/antagonists, progesterone receptor agonists/antagonists, steroid biosynthesis inhibitors, gonadotropin and thyroid modulators either directly or indirectly by altering the hypothalamic-pituitary-gonadal or -thyroidal axes, and prolactin modulators through neuroendocrine pathways.

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 or negative effect on the estrogen, androgen or thyroid hormonal systems.

Although peer review of the intact adult male assay was performed on an individual basis (i.e., its strengths and limitations evaluated as a stand alone assay), it is noted that this assay along with a number of other in vitro and in vivo assays will potentially constitute a battery of complementary screening assays. A weight-of-evidence approach is also expected to be used among assays within the Tier-1 battery to determine whether a chemical substance has a positive or negative effect on the estrogen, androgen or thyroid hormonal systems. Peer review of the EPA's recommendations for the Tier-1 battery will be done at a later date by the FIFRA Scientific Advisory Panel (SAP).

The purpose of this peer review was to review and comment on the intact adult male screening assay for use within the EDSP to detect various mechanisms of action, especially androgen receptor agonists/antagonists, steroid biosynthesis inhibitors, gonadotropin and thyroid modulators either directly or indirectly through intact HPG or HPT axes. 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 15-day intact adult male 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, and the final EPA work product.

1.1 Peer Review Logistics

ERG initiated the peer review for the 15-day intact adult male rat assay on August 30, 2007. ERG held a pre-briefing conference call on September 12, 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. Reviewers submitted all peer review comments to ERG on or before September 24, 2007.

The peer review for the 15-day intact adult male rat assay was initiated on August 30, 2007. A pre-briefing conference call was held September 12, 2007. The purpose of the call was to review the peer review charge and materials, and provide answers to questions or clarification as needed. Reviewers submitted all peer review comments on or before September 24, 2007.

1.2 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.

- ***George Daston, Ph.D.,*** is Research Fellow at Miami Valley Laboratories, The Proctor & Gamble Company, Cincinnati, OH. He has conducted research in the areas of developmental biology; teratology and toxicology, especially mechanisms of normal and abnormal development; nutrient-toxicant interactions; in vitro alternatives in teratology and toxicology; functional teratology; fluid balance in development; and risk assessment. A sampling of his professional activities include, Chair (2006), Task Force for Identifying Refinement and Reduction Strategies for Reproductive Toxicity Testing for the European Centre for the Validation of Alternative Methods; Chair (2002), ICCVAM Evaluation of In Vitro Test Methods for Detecting Potential Endocrine Disruptors for the National Institute of Environmental Health Sciences; President (1999-2000) of the Teratology Society; Committee on Developmental Toxicology (1997-2000) for the National Academy of Sciences/ National Research Council; Endocrine Disrupter Screening and Testing Advisory Committee (1996-1998) for U.S. EPA; and President (1994-1995), Reproductive and Developmental Toxicology Specialty Section of the Society of Toxicology. Dr. Daston is currently Editor in Chief for Birth Defects Research in *Developmental and Reproductive Toxicology*. A few professional journals in which he has published research articles include, *Environmental Health Perspectives*, *Teratology*, *Toxicological Sciences*, and *Reproductive Toxicology*.

- ***Richard Dickerson, Ph.D., DABT***, is an Associate Professor in the Department of Pharmacology and Neuroscience and the Department of Environmental Toxicology at the Texas Tech University Health Sciences Center, Lubbock, TX. He has developed graduate level courses at Texas Tech in chemodynamics, endocrine disruptors, and mechanistic toxicology. Some of the grant-funded research he has performed includes, "Ecological risk assessment of estrogenic and antiestrogenic effects in wildlife exposed to environmental chemicals," "Evaluation of developmental, immunologic and reproductive effects of polychlorinated dibenzo-p-dioxins and dibenzofurans through in vitro assays," "Quantitation of organochlorine residues in human body fat and their effect on MCF-7 growth rate and steroid receptor binding activity," and "Reproductive and developmental

toxicity of TCDD." Dr. Richardson serves on the European Union Endocrine Disruptor Working Group and has Co-Chaired an International Conference held at Kiawah Island on Processes and Principles for Evaluating Endocrine Disruption in Wildlife (March 1996), as well as a two-day symposium held at the national meeting of the Society for Environmental Toxicology and Chemistry in Washington, DC on Endocrine Disruption (November 1996). He has published in several peer-reviewed scientific journals including, *Chemosphere*, *Environmental Toxicology and Chemistry*, *Journal of Toxicology and Environmental Health*, and *Toxicology and Applied Pharmacology*.

- ***Kevin Gaido, Ph.D.,*** is Senior Investigator and Director of the Center for Integrated Genomics at The Hamner Institutes for Health Sciences (formerly the CIIT Centers for Health Research), Research Triangle Park, NC. His current research projects include "Mechanism of phthalate induced testicular toxicity," "Spatial gene dynamics in the fetal male urogenital tract," and "Assessing the impact of chemical exposure on reproductive development," as well as others. Dr. Gaido has served on several committees including, the NIEHS Centers for Environmental Health Sciences Review Committee (2006), and the NIH Clinical Endocrinology and Reproduction (ICER) study section (2005 – 2006). He was an expert consultant for the Center for the Evaluation of Risks to Human Reproduction, a review member of the ICCVAM Endocrine Disruptor Panel, and a subteam member of the Chemical Manufacturer Association's Endocrine Disruptor Test Validation and Standardization. He has published journal articles in *Endocrinology*, *Environmental Health Perspective*, *Reproductive Toxicology*, and *Toxicological Applied Pharmacology* to name a few.

- ***Richard Sharpe, Ph.D.***, is a Professor in the MRC Human Reproductive Sciences Unit of the College of Medicine and Veterinary Medicine at the University of Edinburgh, Scotland, UK. He has over 30 years of experience conducting research in the areas of biochemistry and molecular biology of the development and function of the testis and male reproductive tract, the effects of environmental chemicals and lifestyle factors on testicular and reproductive tract development, endocrinology, fetal/neonatal determinants of adult reproductive health and function, and male reproductive toxicology. From 2000 – 2006, Dr. Sharpe was a member of the Editorial Board of *The Journal of Endocrinology,*

and from 2002 – 2006 he was a member of the Veterinary Medicines Directorate sub-group on hormones and their use in growth promotion. He has also been a member of the Royal Society Working Group that reported on endocrine disrupting chemicals (June 2000), and a member of the COT/Food Standards Agency working group on phytoestrogens (report 'Phytoestrogens and health' published Summer 2003)**.** His numerous research articles have been published in professional journals such as, *Animal Reproduction, Environmental Health Perspectives, Journal of Clinical Endocrinology and Metabolism,Journal of Endocrinology, and Toxicological Sciences*.

- ***R. Thomas Zoeller, Ph.D.***, is a Professor and Chair of the Department of Biology at the University of Massachusetts, Amherst, MA. He conducts research to explore the molecular mechanisms of thyroid hormone action in the developing brain, and the consequences of disruption by thyroid disease or environmental chemicals. His professional affiliations include the American Association for the Advancement of Science, the Endocrine Society, and the Society for Neuroscience. He currently serves on the Editorial Board for *Environmental Toxicology and Pharmacology*, and *Endocrinology.* Dr. Zoeller was previously a Standing Member on the U.S. EPA Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC), Screening and Testing Workgroup (1997-1998). He has organized professional meetings including the 27th New England Endocrinology Conference in Amherst, MA (September 2002), as well as the 23rd New England Endocrinology Conference in Amherst, MA (September, 1995). In September 2000, he served as Session Chair for Endocrine Disruptors at the18th International Neurotoxicology Conference held in Colorado Springs, CO. His published research papers appear in refereed journals including, *Critical Reviews in Toxicology, Endocrinology, Environmental Health Perspectives, Molecular and Cellular Endocrinology, and Neurotoxicology and Teratology.*

2.0 PEER REVIEW COMMENTS ORGANIZED BY CHARGE QUESTION

Peer review comments received for the 15-day intact adult male rat 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.

2.1 Overall General Comments

General comments provided by several reviewers are summarized below.

Richard Sharpe: I have ordered my comments below according to the questions posed to reviewers. However, my placing of some comments is rather arbitrary as, in several instances, there is overlap or uncertainty in my mind as to which, if any, of the questions posed they address.

Tom Zoeller: *Introduction*
Section 408(p) of the Federal Food Drug and Cosmetic Act (FFDCA) requires the U.S. Environmental Protection Agency (EPA) to: *develop a screening program, using appropriate validated test systems and other scientifically relevant information, to determine whether certain substances may have an effect in humans that is similar to an effect produced by a naturally occurring estrogen, or other such endocrine effect as the Administrator may designate [U.S.C. a(p)]*. The 15-day intact adult male rat assay as an alternate component of the Tier-1 screening battery was recommended by the EDSTAC committee and has been developed by industry in the intervening years. The current document represents a considerable amount of effort focused on evaluating the ability of this assay to identify chemicals that interfere with the androgen and thyroid systems. In general, an environmental endocrine disruptor is defined as *an exogenous agent that interferes with the synthesis, secretion, transport, binding, action or elimination of natural hormones in the body that are responsible for the maintenance of homeostasis, reproduction, development, and/or behavior*.

In general, this is an ambitious project that was not managed by EPA in a manner required to achieve the stated goals. This is unfortunate. There are four categories of

weaknesses, each of which was preventable. These include a) lack of performance standards and criteria for RIAs, b) failure to develop a logical framework in which to interpret the results *a priori*, c) failure to carefully control contents of the feed and determine the degree to which this affects the performance of the assay, d) failure to carefully inspect the data generated. Each of these categories is discussed in greater detail below. However, not all of these categories fit neatly into the charge questions; therefore, I will discuss these in greater detail here.

Performance standards and criteria of the RIAs. The RIA data provided in this document show a great deal of variability in hormone levels of the control animals across laboratories. However, it is not possible to identify the source of this variation as being technical or biological because the types of studies required to separate these two sources of variation were not performed. Specifically, the EPA should develop and distribute, or should contract to develop and distribute, the quality control standards to all laboratories performing RIAs in the commission of the EDSP. These centralized standards would greatly decrease the variance across laboratories and would enhance the reliability of the assays. In addition, the three laboratories used different commercial kits for the various RIAs and EPA did not require that the RIAs were validated (in the case of heterologous assays) or that the QC was performed as described by the kit manufacturer or that the performance fell within the range defined by the manufacturer. There is no question that these problems can account for a great deal of variability in the RIA results, and that a minimal amount of thought and effort by the EPA at the beginning of this project could have prevented it. It must be remembered that RIAs have been in use for nearly 50 years, and methods for validating assays and standardizing them across laboratories have been very well developed.

Because of these technical problems, the degree of biological variability in hormone levels and effects of treatments on hormone levels, cannot be ascertained. Certainly, some of the variability observed in this exercise is related to biological variability. One can imagine a number of differences among housing conditions that could account for this. For example, the feed and animal housing in use in the EDSP was not well controlled. We know that there is much greater variability in the contents of the feed

than appears on certificates from the suppliers (1, 2). Differences in the amount of isoflavones in our experiments can make at least a 50% differences in the concentration of total T_4 in serum. Important constituents include not only isoflavones that can act as estrogens and thyroid peroxidase inhibitors, but also iodine, which can greatly influence thyroid function. The EPA made two logical mistakes in the way they present the criteria for the feed. The first paradox is that they argue that 15 days of a specific feed is not long enough to have significant impact on hormone levels or on the response to treatments (without supporting evidence). However, if this is true, then the feed the animals were provided prior to the beginning of the experiment is more likely to have an impact on the experiment, but this is not specified. Controlling the components of the feed will doubtlessly be difficult. However, for EPA to state that, "Certified animal feed will be used, guaranteed by the manufacturer to meet specified nutritional requirements. Analysis will include ensuring that heavy metals, pesticides, and phytoestrogens (e.g., genistein, daidzein, and glycitein) are not present at concentrations that would be expected to affect the outcome of the study", (Appendix C, page 6 of 21) provides no guidance to a laboratory trying to perform this assay to the best of their ability. EPA has not cited information about the effects of phytoestrogens in the feed and the consequences on "the outcome of the study".

Failure to develop a logical framework in which to interpret the results a priori. The EPA document describes in the introductory material (page 4, *Test Development*) that detailed review papers are used as the basis of the test. However, this does not appear to be the case. The review material used do not provide the EPA with a specific framework in which to predict the kinds of effects that would be observed in the 15-day adult male assay. A case in point is the affect of Linuron on the HPT axis. The data presented in this document show that Linuron can produce a significant (and robust) decrease in serum total T_4, but that the thyroid gland and serum TSH is only slightly – or not – affected. Therefore, the serum T_4 levels are considered to be uninformative. This interpretation is supported by the observation that many (27) of the 29 chemicals evaluated in this assay can also cause a decrease in serum total T_4.

This is a highly uninsightful interpretation and reflects that lack of forethought put into the interpretation of possible results. First, it is illogical to base an interpretation on the proportion of chemicals that reduce serum total T_4 in a series of "prevalidation" studies. These chemicals were selected because of preliminary evidence that they are endocrine disrupting compounds. Might they considered a non random sample of chemicals? Second, because the EPA failed to develop endpoints of thyroid hormone action in the 15-day intact adult male assay, the assay itself is asymmetric; that is, there are endpoints of androgen action (organ weight and histopathology), but not of thyroid hormone action. Thus, the assay itself is capable of identifying an antiandrogen that causes a reduction in serum testosterone but does not increase LH, but is not capable of identifying an anti-thyroid agent similarly. The EPA's current interpretation would likely eliminate PCBs as anti-thyroid agents. Although some studies have shown that PCBs can cause an increase in serum TSH, many show that PCBs do not increase TSH levels. Thus, this profile would look like the effects of Linuron and would be ignored. The EPA authors do not explain why two chemicals (Linuron and Phenobarbital) that act by the same mechanism (increase liver clearance of T_4) can have two different effects on serum TSH. To what extent must TSH levels be increased before there are measurable changes in thyroid weight and histopathology? These issues should have been discussed prior to the commission of this assay for inter-laboratory validation and potential solutions identified.

Failure to carefully control contents of the feed and determine the degree to which this affects the performance of the assay. To be sure, this is a difficult task. NIEHS recently sponsored a workshop on animal feed in EDC research and included manufacturers of animal feed. This EPA document ignores the importance of this issue except to state that the level of phytoestrogens should be below that "expected" to interfere with the performance of the assay. In addition, many of these isoflavones inhibit thyroperoxidase and, in our lab, the presence/absence of soy protein in the feed can alter thyroid hormone levels very significantly. Thus, different diets will interact in this assay in a way that increases the biological variability.

Failure to carefully inspect the data generated. Table one [referenced as 2-1 in this report] is a compilation of mean±SEM for testosterone, LH, T_4, T_3, and TSH. These data

were recruited from the individual reports of the 3 laboratories. Highlighted are data cells that contain exactly the same SEM value (to 3 decimal places). For example, T_3 levels in the Linuron-treated groups (0, 50, 100, mg/kg) reported in the WIL report have an SEM of 2.859. Moreover, this value in the Phenobarbital treatment groups is exactly the same. It would appear to be highly unlikely that the standard error of the mean, with 15 animals/treatment group, is exactly the same in all of these groups.

The problems outlined above and described below render this inter-laboratory exercise incapable of being interpreted. It is difficult not to conclude that EPA has not lived up to their charge to validate this assay and the produce a credible document.

2.2 Comments on the Clarity of the Stated Purpose of the Assay

George Daston: In order to provide the reader with an understanding of the purpose of the assay, it is necessary first to provide the context in which it will be used. The summary report does a good job of explaining the legislative mandate for endocrine screening, the tiered approach that EPA has decided to take, and the aspects of the screening tier that are germane to the development of the adult male assay. The only niggling issue that I had with the presentation of regulatory context is the statement that the legislative mandate is part of the Federal Food Drug and Cosmetic Act (FFDCA). It has been explained to me that this is technically correct; however, most of us consider the endocrine disrupter screening program to be a mandate of the Food Quality Protection Act. While I now understand that the FQPA made modifications to both FIFRA and FFDCA, this point escaped me when I first read the report. This confusion is compounded by language on line 13, p. 1 "Subsequent to passage of the Act in 1996", with "Act" referring to FFDCA, a law that was passed more then 90 years previously. It would be an easy fix to add a phrase indicating that the regulatory statute is FFDCA *as modified by the Food Quality Protection Act of 1996*.

The report's interpretation of validation of alternative tests under ICCVAM has a few inaccuracies that should be corrected. These have to do with the interpretation that the validation process was intended specifically for in vitro replacements of in vivo assays (p.

4, line 3, line 45; p. 49, lines 8-9). This is not the intention of the ICCVAM criteria. The criteria are intended to assess whether any assay -- in vivo, in vitro, in silico – is sufficiently robust to serve as an alternative to an existing test method that has regulatory acceptance. ICCVAM has reviewed and accepted in vivo methods as alternatives, including the up-down method for acute toxicity and the local lymph node assay for contact allergy. I don't agree that the ICCVAM criteria represents a "fundamental problem confronting the EPA" as is stated on lines 3-4, p. 4. The major difference between the validation for the endocrine assays and that of other assays is the absence of a gold-standard assay with a large database against which to compare results. This latter problem is the one that the report tries to grapple with, and I agree that it is a legitimate issue. For the sake of clarity in the organization of the report, it would be much preferable to scrap the spurious argument that the validation process is designed for in vitro tests and to acknowledge that because the endocrine screening assays aren't replacing a specific test method some flexibility will be required in how the validity of the new test methods are interpreted.

Given that the purpose of the assay is to identify specific modes of endocrine toxicity, I believe that the correct approach is to validate the performance of the assay using a set of compounds for which the modes of action have been generally agreed upon through the development of a large data set in the literature. The report tries to do this, but it would be much easier to follow if this were presented as the context of providing a standard for validation. For example, one would classify linuron as having anti-androgenic activity or Phenobarbital as having thyrotoxic activity based on a critical review of the literature. This review includes the verification that this activity has adverse consequences on the male reproductive system or thyroid, respectively, in a toxicity study that conforms to regulatory guidelines. This approach would satisfy the validation criteria and provide a basis for making calculations of assay performance (e.g., concordance, sensitivity, specificity, etc.).

One of challenges for the report's authors is to clearly present information on how to interpret a test as complicated as this one. The report could be better in this respect. I would especially like to see a section that provides criteria for interpretation. There are

perhaps a dozen modes of action that this assay was designed to detect: thyroid disturbances, androgen receptor agonists and antagonists, estrogen receptor agonists and antagonists, progesterone receptor agonists and antagonists, inhibitors or enhancers of steroid synthesis, dopamine agonists and antagonists (assessed via prolactin modulation) and other modes that perturb the pituitary response within the hypothalamic-pituitary-gonadal axis or the hypothalamic-pituitary-thyroidal axis. Ideally what I would like to see is a short description of which assay endpoints would be changed in order to categorize something as an androgen antagonist, steroid synthesis inhibitor, etc. This could be done in the text, or as a flow chart. If it is necessary to include information from other tier 1 screening assay, that's fine, as it appears from Table 3 that this test would be performed as part of a battery. I realize that there will be a need to modify these interpretations as more data become available for this test method. However, the endpoints in the assay were selected based on solid mechanistic understanding that by measuring them it would be possible both to detect certain modes of action and rule out others. The developers of these assays have been stating such interpretations since first publishing on these tests in the '90s (O'Connor et al., 1996; Cook et al., 1997). It would be useful to have short summaries here. The appropriate place would be at the end of section 3.

The intact male assay has been used extensively by a number of industry labs. The data from these labs is summarized nicely in the Prevalidation section of the report. However, the industry groups appear to be using the assay for a broader range of modes of action than EPA evaluated in its validation study. Table 4 lists such modes of action as progesterone receptor agonism/antagonism, dopamine receptor agonism/antagonism, that were beyond the scope of the validation program. It wasn't clear to me in reading the report whether this was simply due to the limited scope of the initial validation, or if EPA intends to scale back the purpose of the study. I would like to see this point addressed specifically, in section 2.1.

Richard Dickerson: First, the purpose of the report and assay should be stated much closer to the beginning of the report than on page 6. The assay and its validation are the focus of the report, not the history of why it is needed. I suggest placing section 1.6

(Purpose of the ISP) as 1.1 followed by the purpose of the assay (2.1) followed by the remainder of the introduction. This allows those individuals familiar with EDSTAC to focus on the purpose of the report and assay and perhaps skip the historical background. Another option might be to include an executive summary following the cover sheet that summarizes the purpose of the report, the purpose of the assay and the conclusions derived from the results of the interlaboratory validation. If the target audience is the decision makers, putting the bottom line up front provides greater assurance that they will get the message.

Second, the purpose of the assay could be more directly and clearly stated. It is stated in passive voice rather than active, and begins with a reference to other publications. It is more effective to state "The purpose of the 15-day intact adult male rat assay is to detect compounds or mixtures that alter the HPE, HPA and HPT through the most probable MOA." The list of MOAs can follow along with the endpoints measured. A brief description of assay methodology can be included.

Third, if the third paragraph (beginning on line 8 of page 7) is to be included with the purpose of the assay, consider adding a reference to Table 4. This allows a quick comparison of assay capabilities.

A somewhat related comment is that a discussion of progesterone and RU486 was not included in section 3.1.1 Positive Test Chemicals. This should be added.

Kevin Gaido: The stated purpose of the assay, as an alternative to the female pubertal assay to detect chemicals that interfere with androgen or thyroid function, or through the HPG axis is clearly stated. The goal is to develop a relatively quick, reliable screening assay that will be part of a comprehensive battery of tests for endocrine active chemicals.

Richard Sharpe: The background information and discussion provided give a clear view of what the assay is intending to achieve and why it has (most of) its component parts. It is a Tier-1 assay and, as such, its priority is to maximize the detection of endocrine active compounds whilst minimizing false negatives. The use of multiple

endpoints is designed to ensure this. Its particular strength, discussed in more detail later, is that it should sidestep issues related to hormone homeostasis, which is always likely to be the main confounder in an assay such as this which uses an intact animal with normally functioning homeostatic hormone systems.

I found the information on the purpose of the assay and its background to be clearly presented, easily understandable and to make commonsense. It should perhaps emphasize that the assay is not intended to be definitive, as this is important when considering results from individual laboratories (for example, in the inter-laboratory comparison) in which inconsistency in results may occur, but in which the assay achieves its primary objective.

Tom Zoeller: The stated purpose of the assay is perfectly clear. A point of confusion though is the relationship between validation of individual assays and validation of the battery. Tier-1 and Tier-2 batteries are complex, and for them to be informative as envisioned, each of the component assays must be reliable and their interpretation must be guided within the context of the tier itself. However, the discussion in the document does not clarify the relationship between validation of the 15-day adult male assay and the Tier-1 battery itself.

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

George Daston: The primary purpose of the study, as described in section 4.1, was to evaluate the reliability and transferability of the newly developed standard protocol, and to a lesser extent to continue to assess assay relevance. Given the primary purpose of the study, I believe that the right data are emphasized in section 5 of the report. The authors of the report stayed focused on the goals of transferability, reliability, and adherence to protocol.

There are comparisons to historical control data, particularly for body and organ weight, that might be interpreted differently if additional historical control data were considered (Table 9 and the accompanying text). The historical control data appear to be limited to

28 studies using a similar study design and compiled by O'Connor et al in 2002. Many of the studies in O'Connor's paper were several years old at the time, and are now more than 10 years old. There is a constant, subtle drift in body and organ weights over time such that the older data may not be as relevant. Furthermore, it isn't possible to know whether subtle differences in housing conditions or husbandry in the various labs produces variability in relative organ weights. Therefore, it would be useful to include historical control data from each of the three labs for this species and strain of rat. It is likely that they have data for control body weights for SD rats from 10-12 weeks of age because this is within the age range of animals used for subchronic and reproductive toxicity studies. The age-range is a little young for organ weight data from 91-day subchronic studies, but relative organ weight (organ/body weight) may be informative.

Richard Dickerson: The data from each of the laboratories was presented clearly and factually. The CVs both within a laboratory and between the three test laboratories were adequately analyzed and discussed. The largest area of concern, i.e. the large variability in certain hormones, was identified and thoroughly discussed. The results of each of the assays were also correctly interpreted and sources of error identified. The sets of results from the linuron exposure and the phenobarbital exposure were presented in the same format. However, analysis of the results and interpretation were more thorough for linuron than phenobarbital.

For the linuron exposure study, unacceptable variability in the results of the assays for prolactin (PRL), testosterone (T), dihydrotestosterone (DHT) and at the highest dose thyroxin (T4) occurred. It is also of concern that for some of these endpoints either no change was observed when historical data had reported an effect or an effect was observed when historical data suggested no change. However, there were no instances where the direction of change was opposite to those previously reported. For many of the androgenic endpoints, one or more laboratories failed to detect an effect although previous studies had found decreases. It is interesting that the results obtained by Charles River more frequently matched the historical trend whereas the results for RTI did not despite the fact that RTI performed hormonal analyses for both.

For the phenobarbital exposure, unacceptable variability in the assays for T, DHT and PRL occurred at all dose levels. However, the thyrodogenic endpoints and the liver weight changes were significant as predicted by other studies.

Kevin Gaido: The summary statement provides a clear and comprehensive interpretation of the data. A detailed comparison of the results from each laboratory together with historical data is provided. To allow for sufficient interpretation of the results.

Richard Sharpe: Considerable data on this assay has been collected involving several laboratories and a large number of compounds with a wide variety of mechanisms of action (MOA). The evidence presented in reports and publications, primarily those by O'Connor *et al*, substantiate the view that this assay is fit for purpose. An important point that is made repeatedly, and which cannot be overemphasized, is that this assay intentionally uses multiple endpoints in order that it may more readily identify compounds with weak activity or with a profile of activity that does not fit within expected boundaries (for example a compound that exhibits both anti-androgenic and anti-thyroidal activity). The other purpose of the multiple endpoints is to provide preliminary information on the potential MOA, which may then guide decisions about subsequent testing in Tier-2. However, in my opinion, the main importance of the inclusion of multiple endpoints in this Tier-1 assay is to maximize the likelihood of detection of endocrine active chemicals whilst minimizing the chance of false negatives.

The interpretation of the results obtained using this assay in the different laboratories, including the inter-laboratory validation exercise, are rational and fit with current understanding of how the various endocrine systems operate within the body. Every aspect of the data has been evaluated in terms of its robustness, its reproducibility, sensitivity of detection and consistency with other results in the same assay from the same laboratory or with results from other laboratories. There are some minor issues in relation to homeostatic changes (see my comments to question 5) and there are issues in relation to interpretation of weight changes for the epididymis, but these do not affect the overall conclusion that the assay is robust, but with some limitations. In making these comments, I base them very much on the pre-validation studies that involved extensive

testing of a wide range of compounds rather than on the inter-laboratory validation exercise. If my evaluation was based on the latter alone, I would be less enthusiastic about the utility of the assay and again I discuss this further in relation to question 5 below.

Although the O'Connor studies using chemicals with well characterized anti-androgenic activity via one or more mechanisms (flutamide, ketoconazole and finasteride) are highly convincing in this assay, as would be expected, interpretation of results for compounds with less dramatic activity might be more equivocal if the only results available were from the present assay. One such example is results with vinclozolin, which even at 150mg/kg, only resulted in a significant reduction in epididymal weight with no significant effects on relative seminal vesicle or prostate weight and only a significant elevation in LH levels with no change in testosterone. Nevertheless, within the stated aims of the assay, this compound would still be flagged up for further study. Similar results to vinclozolin were obtained for linuron in the prevalidation studies and inter-laboratory validation exercise and, if changes in thyroid weight and thyroid hormone levels are ignored, then it is only the change in epididymal weight at higher doses of linuron exposure that would flag this compound up as a potential anti-androgen. These particular comparisons also illustrate the limitations of the assay in terms of identifying the MOA, as I am not sure that I would be able to identify an MOA based on the profile obtained for linuron. It may therefore not always be possible to definitively design Tier-2 investigations based on an MOA discerned from the Tier-1 screen using this assay.

Tom Zoeller: The manuscript clearly describes the logic used to interpret the data provided by the 3 laboratories. The methods employed and the endpoints collected are clear. The EPA document, and the individual reports from RTI, WIL and Charles River, indicates that because the RIAs are so variable both within and between laboratories, the hormone levels are to be used for supportive evidence for a role of a chemical as an endocrine disruptor (androgen or thyroid), but that body and organ weight and histopathology should represent primary data. Thus, the endpoints captured including body and organ weight and histopathology (thyroid, testes, epididymides), provide primary information about the toxicity of a chemical and the MOA as an endocrine

disruptor. There are two problems with this logic. First, the tissues employed as endpoints of androgen and thyroid disruption represent endpoints of androgen action (e.g., epididymus, seminal vesicles), but there are no endpoints of thyroid hormone action that would be equivalent to epididymus or seminal vesicles. Thus, chemicals like linuron that can reduce circulating levels of thyroid hormone without affecting (or perhaps even lowering) serum TSH may not produce an effect on the thyroid gland itself (through elevated TSH) and will therefore be ignored. Thus, there is a fundamental flaw in the endpoints designed for capture in this assay. Second, although a considerable problem is that of the high variability in hormone levels both within and across laboratories, there may be a solution to this problem (see below). In the absence of providing reliable data for hormone levels, this and the other in vivo assays will be severely compromised.

The relationship between body weight reductions produced by toxicity or by caloric restriction is a complex one and the background information provided is interesting and important. Briefly, these data show to what extent total body weight must be reduced (caused by caloric restriction) before impacting the weight of the various organs or hormone levels. This information is used in the interpretation of the data arising from toxicant treatments by assuming that the relationship between total body weight and organ weight will hold for all toxicants. Caloric restriction is known to produce a significant and potent reduction in serum thyroid hormone levels, which can be blocked by placing lesions in the hippocampus (3, 4). Thus, the fasting-induced reduction in thyroid function is mediated by the central nervous system. In addition, this effect also involves the type 2 deiodinase (5, 6). Therefore, the effect of caloric restriction on the HPT axis is centrally mediated and may respond to toxicants in ways that do not simply duplicate caloric restriction. Perhaps changes in the use of specific metabolic fuels (fat, protein, carbohydrate) can elicit this response in the absence of large changes in body weight. In contrast, perhaps some chemicals can block this effect regardless of body weight changes? Although somewhat speculative, this hypothesis is clearly plausible and the simple assumption that body weight will always be related to organ weight in a particular way seems both unnecessary and dangerous.

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

George Daston: I believe that the biological and toxicological relevance of the assay is well described in section 3.1. I believe that this assay, as well as the pubertal male and pubertal female assays being evaluated, has the potential to provide the most reliable and comprehensive information for the weight-of-evidence determination described in section 1.4. The use of an intact animal model provides the opportunity to assess multiple endocrine processes, both alone and in integration with the hypothalamic-pituitary axes that control thyroid and gonadal function. The ability to measure multiple modes of action in a single assay provides the opportunity to obtain a lot of information from a relatively small number of animals, vs. running separate tests for each mode of action. The intactness of the hypothalamic-pituitary-gonadal and hypothalamic-pituitary-thyroidal axes makes the model biologically relevant, as these axes act in concert in the organism that we wish to model for the purposes of hazard and risk assessment, the human. The model is toxicologically relevant because the responses in an intact system, which also has homeostatic mechanisms, is likely to be much more concordant with the results of more definitive toxicity tests.

Richard Dickerson: In terms of biological relevance, the assay endpoints reflect measures of the integrity of the hypothalamic-pituitary- androgen (HPA) and -thyroid (HPT) axes. These include changes in tissue weight, histology, and circulating hormone levels. Other assays relevant to the androgen axis might include rate of sperm production, sperm motility, and ability to undergo the acrosome reaction. However, the length of the cycle for sperm production greatly exceeds the 15-day period of chemical exposure used in this assay. Other measures of reproductive capacity also require much longer times of exposure than used for this assay. The endpoints used for the HPT axis are also the most appropriate for the length of the assay.

In terms of toxicologic relevance, the endpoints selected for the 15-day Adult Male Rat Assay are appropriate for several reasons. First, they reflect biologically relevant endpoints as discussed above. Second, previous studies using known androgen receptor agonists and antagonists demonstrate these endpoints are altered by exposure to methyl

testosterone, vinclozolin, flutamide, p,p'-DDE and other AR agonist/antagonists. Finally, the endpoints are relevant because competent investigators, whether from industry, contract laboratories or academia are capable of measuring them in a consistent manner.

Kevin Gaido: As stated above, the assay was designed to detect chemicals that interfere with androgen or thyroid function or with the HPG axis. While of little biological relevance, this assay is highly relevant for toxicological screening for endocrine active chemicals.

Richard Sharpe: From the pre-validation exercise, a strong foundation has been laid for evaluation and interpretation of results in the assay for compounds for which no information exists about their potential hormone activity. The multiple endpoints of the assay and its relative simplicity mean that its continued application will lead to a progressive ability to categorize chemicals into classes based on their activity profile, even when it is not possible to define a clear MOA. As the profile database expands, so the toxicological utility and predictability of the test is likely to expand also. Because the test uses an intact, adult animal, then compounds may affect target organs or hormone levels via pathways that are unrelated to endocrine disruption *per se*, for example effects on food intake/metabolism that leads secondarily to such changes. This is the 'real world', and it is a strength of the assay that it can integrate such 'biological' effects, though a further reality is that it may be difficult to disentangle such effects from primary endocrine effects in some circumstances (see Q5 below).

Tom Zoeller: Section 3.1 discusses the relevance of the bioassay. This section begins with statements about how "Numerous EACs (one negative and 28 positive test chemicals)…" have been tested in the 15-day intact adult male assay, but the data are not presented nor are they fully referenced. In addition, an examination of Table 4 lists these chemicals with a very cursory description of their MOA. For example, the document states that, "Thus, throughout prevalidation, the intact adult male assay has been run with 29 different test chemicals at various times in six different laboratories (four chemical industry laboratories and two different contract research organizations, or CRO laboratories). In some instances the same chemicals were tested in more than one

laboratory at different times as shown in Table 4." This is a very misleading statement that is not supported by the information presented in Table 4 nor is it supported by the discussion in Section 3. Therefore, it undermines the credibility of the current document more than it supports the strength and validity of the 15-day adult male assay.

In addition, it is not clear from the remainder of this section why this information is being presented. This section could have provided a logical basis for the design of the assay. To do so, it would have to review the basic endocrinology of the androgen and thyroid systems to the extent that the biological and toxicological relevance of a 15-day assay performed in the adult male would be supported. However, no such discussion is presented, and the choice of endpoints identified in the 15-day adult male assay lack clear support, which becomes apparent when the data are interpreted. Section 3.1 presents a very cursory review of a variety of chemicals evaluated in a variety of experimental designs. This is a highly confusing section that does not advance arguments in support of the assay. Worse, this section indicates that a variety of chemicals with known endocrine activities were evaluated in the 15-day adult male assay. Thus, it gives the impression that chemicals were defined as having endocrine activity in other kinds of experimental designs and then evaluated in this 15 day assay. If this assay has provided fundamental new information about the endocrine activity of various chemicals, this section does not provide a credible review of this. Overall, this section is a pivotal section that fails to provide a careful review of the literature that is relevant to the 15-day adult male assay, nor does it provide a convincing argument that supports the expectation that this assay could provide information about the biology of the androgen or thyroid system. Indeed, it is difficult to imagine that new insight into these endocrine systems will be generated from such an experimental paradigm.

Finally, the toxicological relevance of this assay also is not made clear in this section. This is not to say that the assay does not have toxicological relevance. Rather, the information required to conclude that this assay has toxicological relevance is simply not presented. Specifically, the document begins with a definition of an "endocrine disruptor". This definition, copied verbatim above, should provide the foundation of section 3.1. That is, if a chemical acts as an EDC (by definition), then one can make

predictions about the endpoints that should reveal this and can be used to identify EDCs. The form of this section is not such that this logic is presented.

Section 3.2 also is an important part of the document that could provide a logical framework for the design and interpretation of the 15-day adult male assay. However, this section is written in such a way that it fails to provide credibility to the overall document. For example, this section introduces the concept of chemically-responsive "fingerprints" that may provide information about the mode of action of a specific chemical. This concept of "fingerprints" turns out to simply mean that if endpoints at different levels of the HPG and HPT axis are evaluated, one may obtain basic information about the site at which a chemical interferes with endocrine activity. As an example, work attributed to O'Connor et al. (1998a and 2000c) is provided. This work apparently showed that, "Correspondingly, the ability of flutamide to block the negative feedback effect of testosterone and DHT at the hypothalamic and pituitary levels resulted in the secretion of gonadotropin-releasing hormone (GnRH) and LH, respectively, and the subsequent production of testosterone by the Leydig cells of the testes. Thus, the chemical-responsive "fingerprint" of an AR antagonist such as flutamide is a decrease in ASG weight and increased serum concentrations of testosterone and LH." This is exactly the kind of logic that should have been presented in section 3.1 and the basis for that logic is basic endocrinology. However, the references that appear to be used to support this statement do not, in fact, provide support. Neither of these citations measured – or even mentioned – GnRH, and one (O'Connor et al., 2000c) does not report studies of flutamide. Moreover, even a cursory knowledge of research on GnRH would have informed the writers of this section that GnRH secretion is a very technically demanding endpoint to capture, and there are not many laboratories with the skill or equipment to perform such studies. Thus, to make a statement such as that cited above without the proper support undermines the credibility of the document in providing a logical framework upon which this 15-day adult male assay is developed.

There are two important points that this section (3.2) illustrates. First, the concept that identifying "fingerprints" of endocrine activity as a novel approach requires that one suspend decades of basic research in endocrinology that informs such an approach.

Essentially, this "fingerprint" simply means that one may infer the site within an endocrine axis that a chemical acts to interfere with the system by simultaneously capturing endpoints at different levels within the axis (e.g., gonadal and pituitary hormones). It is reasonable that the 15-day adult male assay for EDCs be placed within an endocrinological context to have credibility both as an individual assay and as a component of tier-1 screens. However, the endocrinological context is not provided in this document and the writers appear to be unaware of this context. This undermines the presentation of the data and its interpretation later in the document.

2.5 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

Richard Dickerson: My comments are based on the protocol appended as C since section 4.0 states it is the final, standardized protocol.

Kevin Gaido: The protocol is clear and comprehensive. The objective is clearly stated and sufficient detail is presented to allow a laboratory with the appropriate expertise to conduct the assay and accurately analyze and report the results.

Richard Sharpe: Insofar as I feel able to judge (as a scientist running an academic research laboratory), the protocol provided is clearly laid out, understandable and sufficiently detailed to enable an appropriately experienced laboratory to run, complete, evaluate and report results using this assay. There are no major deficits in the protocol that I have spotted, but there are some aspects that could potentially cause confusion. Chief amongst these is the inclusion of hormone assays for FSH, estradiol and to a lesser extent DHT. Although FSH is a key reproductive hormone in the male, its inclusion in the present assay is not especially informative and is not clearly defined. I am uncertain how easy it will prove to interpret treatment-induced changes in FSH levels in the context of the aims of the assay (see comments to Q6), and this may cause confusion to future users of the assay unless its role and significance are better defined (eg its main purpose is to support data for LH to highlight compounds that suppress hypothalamic-pituitary

function). The same comments apply to estradiol, as I am unconvinced that our understanding about the role, regulation and significance of changes in estradiol levels in adult male rats is established sufficiently to enable its use in the assay in an informative, as opposed to a confusing/confounding, way. I am not convinced that DHT measurement adds any value to the assay (see comments to Q6).

Tom Zoeller: This question appears to refer to Appendix C. Thus, the answers below are focused on this section.

2.5.1 Comprehend the Objective

George Daston: Most of the information regarding the purpose of the assay is in the Introduction. I found that this section had some of the same clarity problems as the report. Specifically, it is not clear whether the purpose of this standardized protocol is to conduct an assay that is capable of detecting all of the modes of action listed in the first paragraph (p. 4 of the protocol), just estrogen, androgen, or thyroid-related modes, as is implied in the second paragraph, or the list of modes described in the third paragraph: AR agonists and antagonists, steroid biosynthesis inhibitors, gonadotropin and thyroid modulators. It would be easier for the lab to understand the nature of the work if two of these three were eliminated from the introduction to the protocol.

Richard Dickerson: Under objective the statement <enter the specific purpose of the assay> appears. It is therefore not possible to evaluate the clarity and conciseness of the objective. The section on personnel is also incomplete.

Tom Zoeller: The objective of the 15-day intact adult male assay is to contribute to the first tier of screens for EDCs. Thus, it is intended to identify *new* chemicals (i.e., chemicals for which little information is available) that interfere with estrogen, androgen or thyroid activity. This much is clear.

2.5.2 Conduct the Assay

George Daston: The right information is present in the protocol, as evidenced by the fact that none of the three labs had any serious deviations from protocol.

Richard Dickerson: The instructions on how to conduct the assay are complete and clear. However, certain areas are troublesome. First, the dosing solutions are made in 0.25% methylcellulose in water for this assay but are prepared in corn oil for the pubertal male assay. What is the reason for this inconsistency? In addition, many water supplies have measurable amounts of perchlorate. Although these are usually below a level of concern for the general public, it should a mandatory analysis for this assay. If perchlorate is detected, the animals should not receive water from this source. Feed samples should be analyzed for phytoestrogens, all food for a given study should be from the same lot, and it would preferable for all laboratories to use the same food source. A feed low in phytoestrogens would be better than standard rodent chow. In terms of euthanasia, other animals should not be present in the necropsy room when an animal is euthanized or necropsied. A number of studies have demonstrated that when animals in the room when another rat is euthanized or necropsied experience significant increases in corticosterone and prolactin. If all the rats are in the room, the stress hormone levels will be markedly different between the first animal euthanized and the last. Although transporting the animals together minimizes one source of variability, it introduces another source if the animals are all in the necropsy room. The protocol does not specify whether the euthanasia chamber is to be precharged with carbon dioxide gas or if it will be added slowly. The protocol does not specify whether pure carbon dioxide is used or if a mixture of carbon dioxide and oxygen is to be used. A specific technique should be utilized in all studies and it must conform to the most recent AVMA guidelines for euthanasia. In terms of hormone assays, what percentage of samples will be run as true duplicates? In addition, most RIA kits use I125 which has a relatively short half life. Using a fresh kit for some of the samples and an older kit for other samples can introduce variability. The protocol should specify that sufficient kits with the same lot number should be ordered so that all assays for a particular hormone are more consistent. Perhaps a standard sample could be prepared and sent to the laboratories as an additional

QC standard. Last, samples shipped from one laboratory to another require detailed chain of custody documentation and if possible a data logger so that sample temperature and time can be documented. Minimal standards for transit time and temperature set.

Tom Zoeller: The methods described appear to be sufficient to guide an independent laboratory to conduct the assay.

2.5.3 Observe and Measure Prescribed Endpoints

George Daston: The experimental design is very detailed and specific. The procedures seem clear and interpretable.

Richard Dickerson: clear and concise

Tom Zoeller: The information provided is sufficient in most cases. However, as described more completely below, the EPA should provide additional guidance and criteria for helping independent laboratories perform hormone analysis.

2.5.4 Compile and Prepare Data for Statistical Analyses

George Daston: The procedures for data compilation and statistical analysis are clear.

Richard Dickerson: clear and concise but consider specifying statistical software.

Tom Zoeller: The information provided is sufficient.

2.5.5 Report Results

George Daston: The protocol is very clear about the data that should be summarized, even down to the point of prescribing which data should be in tables, which in figures, and how the figures should be drawn. This level of control over data presentation is more than what I am accustomed to seeing. I believe that it is useful to have this level of

control when making head-to-head comparisons of interlaboratory results. However, it may be prudent to remove it should the assay become routine.

The section on interpretation of effects (pp. 15- 18 of the protocol) was surprising to find in the protocol. Given that the primary purpose of the study was to determine reliability and transferability, it would seem to me that this kind of information is more relevant to the study sponsors than the participating labs. I question that it should be in the protocol.

Richard Dickerson: clear and concise.

Tom Zoeller: The information is sufficient.

2.6 Comments on the Strengths and/or Limitations of the Assay

George Daston: The strengths of the assay are nicely laid out in section 2.3 of the report. As noted in my response to question 3, I consider the fact that this is an intact system is a major strength. It should be possible, with a limited number of measurements, to obtain information on a number of modes of endocrine action. The level of information may actually be more than what one can obtain from many definitive toxicity tests. This assay (or its alternates, the pubertal male or pubertal female assay) will provide the greatest weight in weight-of-evidence schemes that will be applied to the tier 1 battery.

I agree with the limitations and challenges given in section 2.3.2. I would add one or two more. First, it is not clear yet whether this assay will have the sensitivity that other assays have, largely because it is an intact model. The fact that there are homeostatic mechanisms in place will tend to blunt – not overcome, but blunt – some of the responses that are being examined. One of the potential strengths of this assay is that some of the potential for homeostasis can be unmasked through the measurement of hormone levels.

The hormone measurements are among the more important aspects of this protocol, especially if one of the goals of the assay is to obtain a mode of action fingerprint. I take

the results of the interlab comparisons of hormone levels to be promising, except for two of the hormones which are probably the least important contributors to the resolving power of the assay, at least for the modes of action it will be applied to in the EPA tier 1 battery.

I found the apparent lack of specificity of T3 and T4 as indicators of thyroid toxicity to be troubling. It will be necessary to develop much more data on negative compounds (i.e., non-endocrine disrupters) in this assay to ensure that changes in thyroid hormone levels can be appropriately interpreted.

Richard Dickerson: Strengths of the assay include ease of conducting the assay and measuring the endpoints, the short duration of exposure, biological relevance of endpoints and robust database.

Limitations of the assay are its inability to determine more downstream effects such as sperm production, motility and fecundity. However, assays that detect these endpoints are more appropriate for Tier-2. The intra- and interlaboratory variability in the hormone assays make it more difficult to detect subtle changes with any degree of significance.

Kevin Gaido: Strengths of this assay include the ability to screen for multiple modes of action in an adult animal. The assay has multiple sensitive endpoints that can be used to help design more definitive Tier-2 testing. Because it is in vivo the assay allows for consideration of absorption, distribution, metabolism, and excretion. The assay is relatively rapid and has been standardized so that it can be performed in any laboratory that has the appropriate expertise and experience.

A weakness of this assay is the necessity of blood hormone measurements. These measurements are highly variable, inconsistent and subject to experimental conditions. Hormone measurements are not routinely done in toxicology studies and many laboratories will not have the appropriate expertise. The inconsistent results across laboratories with linuron suggests that this assay may not be reproducible for weak

androgen receptor agonists. In addition, previously published studies indicate that this assay may not be a sensitive screen for week estrogens.

Richard Sharpe: The main strength of the assay is that it has a strong foundation based on the pre-validation studies using a variety of compounds with different activities, and the use of multiple endpoints that extend beyond organ weights to include evaluation of hormone levels and how their homeostasis may have altered (at one acute time-point). At the same time, the use of multiple endpoints, and in particular hormone concentrations, raises the possibility of sporadic false (chance) results and identification of false positives. Although these may be weeded out by evaluation of the overall result profile for the compound in question, this may not always be possible, but is probably acceptable as no test will ever be 100% perfect. Of more obvious concern is if there are false negatives. In this regard, it is of interest to consider the results for linuron from the inter-laboratory validation exercise in some detail, as based on its profile this compound may have come close to being classed as negative, if data for the thyroid axis were excluded from this analysis (and the presumption is that the thyroid changes are secondary to changes in bodyweight and liver weight). My concern is that identification of linuron as a positive compound in Tier-1 depends very much on its effect on epididymal weight and its classification as an anti-androgen is really not possible from the profile obtained, particularly based on results from two of the laboratories involved in the exercise and on comparison of hormone results with those obtained after Phenobarbital treatment (for which no reproductive axis effects would be expected). With linuron, there were no significant changes in testosterone or LH levels in two of the laboratories and no significant changes in relative prostate and accessory sex organ (ASG) weights in two of the laboratories, so it is not obvious that this compound is an anti-androgen. The elevations in FSH levels and in estradiol levels as measured in at least two out of three laboratories suggest that something is going on (though this profile means nothing to me!) but would not class this compound as an anti-androgen. Throughout the report and throughout the inter-laboratory validation exercise, changes in epididymal weight are viewed as evidence of anti-androgenicity, but I am not convinced that this is a logical conclusion. The epididymis is undoubtedly an androgen target organ but this is not nearly so obvious as for prostate and seminal vesicles (weights) and the weight of the adult

epididymis is probably determined more by the number of sperm that are present, and stored, in the cauda epididymis (reflecting the completeness of spermatogenesis) than androgen effects *per se*. Though anti-androgens can perturb spermatogenesis, the testis is highly resistant to such effects due to the local, intratesticular production of testosterone. Published studies indicate that linuron may have mixed activity which includes direct effects on the Wolffian duct/epididymis as well as some ability to perturb androgen production, although most such studies have been on the male fetus rather than on the adult as in the present assay. My concern is that when the assay is run in the future in one laboratory for a compound such as linuron, but for which there is no pre-existing data, would it end up as a false negative? I think this is probably unlikely but I use this discussion to illustrate this as a potential limitation of the assay. Overall, I consider that there are sufficient, if inconsistent, changes in the results profile for linuron in the inter-laboratory validation exercise for it to be flagged up for further study for its reproductive effects in Tier-2.

Where the strengths versus weaknesses of this assay are very much in the spotlight is when it comes to analysis of the hormone profile. It is a theoretical strength of this assay that it uses an intact animal ("the real world") in which normal, homeostatic endocrine systems are operating. When any one component of an endocrine loop is disturbed, there should be compensation to bring this axis back to normal levels in terms of biological function. As a consequence, for example, there may be suppression of testosterone production by a compound which then triggers increased LH secretion to act on the Leydig cells to bring testosterone levels back to normal. Measurement of testosterone levels after such an adjustment has occurred may not indicate that anything has happened whereas in fact supranormal LH levels are required to maintain the normal level of testosterone. Such a situation is commonly referred to as "compensated Leydig cell failure". In the pre-validation studies, using compounds with pronounced and established anti-androgenicity, such as flutamide and ketocomazole, such changes in the LH-testosterone axis are highly evident, but this is not the case for much weaker anti-androgenic chemicals such as vinclozolin or, as shown in the inter-laboratory validation exercise, for linuron. If a compensation in LH levels in such situations is relatively minor, this may not be easily discernible against the natural background variation in LH and

testosterone levels, which show wide normal fluctuations due to the episodic nature of their secretion. One simple way in which the present assay could be improved to detect such changes would be to determine an LH:testosterone ratio for each individual animal (and then derivation of mean values for the group etc), as this can often indicate that there has been a chronic readjustment of the axis irrespective of what the actual LH and testosterone levels are at any one time in an individual animal; essentially, this ratio is a readout of the current dynamics of the pituitary-Leydig cell axis. This would be a simple refinement to the present study and might enable better identification of such a readjustment in the pituitary-testicular axis, although it is possible that such an adjustment may occur only for a period of time outside of the sampling time used in the assay (see next).

The assay as currently designed involves dosing of animals on day 15 some 2-3 hours prior to euthanisation and sample collection. This is probably a wise choice as it increases the chance of detecting transient effects on hormonal axes that may otherwise not persist to be detectable at a later post-dosing time. The downside of this is that it may detect treatment effects that are relatively trivial, and are sufficiently transient to have no detectable biological consequence or that it may fail to detect effects that are latent (unless these persist to the following day or 'accumulate'). Such considerations prompt me to conclude that the hormone data should be viewed primarily as playing a supporting role for organ weight/histopathology changes (which provide a summation of effects throughout the treatment period) rather than the other way around. Otherwise, I would be forced to conclude that linuron and phenobarbital have a similar, though not identical, MOA as both mildly suppress testosterone (and possibly LH) levels and elevate estradiol levels, whereas only linuron has any suppressive (anti-androgenic) effects on ASG/epididymis.

It is a fact of life that hormone measurements in the same blood sample in different laboratories can yield dramatically different values for absolute hormone levels, even when all the laboratories are using the same assay kit and procedures. This is well illustrated in the present inter-laboratory validation exercise. Whilst such variation can be taken into account by use of a quality control system, as done presently, this never

completely resolves the problem and the bottom line is that it is always difficult to make definitive comparison of absolute hormone levels ***between*** laboratories. There is much more strength when considering changes in absolute hormone levels in different situations ***within*** one laboratory. This is a problem that can be minimized but it will never be completely resolved and the workings of the present in vivo adult rat assay therefore have to take this into account. It is a considerable strength of the adult male assay that it is not reliant on hormone changes *per se*, but on hormone changes in relation to changes in target organ weights.

Some of my reservations about clear identification of linuron as an anti-androgen in the inter-laboratory validation exercise would be removed if absolute ASG weights were used rather than organ weights relative to bodyweight. In this case, linuron would be flagged up as a very clear anti-androgen whereas Phenobarbital would not. However, based on the pre-validation studies by O'Connor using restricted feeding, it was clearly shown that weights of the ASG, other than the epididymis, all declined in parallel with declines in bodyweight. Although these declines were only evident for decreases in bodyweight of 15-25% (again leaving out the thyroid data), this encompasses the magnitude of change in bodyweights for the higher doses of linuron in the inter-laboratory studies. Whilst I understand the basis for using only relative sex accessory organ weight, I wonder if note will be taken of the absolute organ weights when considering the overall profile and classification of any test compound? Correction of organ weights for changes in bodyweight will help minimize identification of false positives, but my concern would be that it may also result in false negatives as might nearly have happened for linuron in at least one of the laboratory studies. I am not certain that it can be concluded definitively that a decrease in bodyweight will always lead to a secondary reduction in ASG weight, irrespective of the mechanism of action that initially precipitated the reduction in bodyweight; it also needs to be remembered that reductions in testosterone levels may itself result in loss of bodyweight/altered body composition which is a potentially confounding effect (though may not be too important in this relatively short assay).

An undoubted strength of the assay in terms of its multiple endpoints is that it also has the potential to identify compounds which have reproductive or thyroid target organ effects but which do not operate primarily through endocrine effects. It is possible that linuron may be a somewhat unclean example of this as it may have direct effects on the epididymis as well as endocrine effects directly on the testis. However, in the context of development of the present Tier-1 assay, it is unclear to me if such compounds would be flagged up for further study if there is no evidence for any endocrine activity

Tom Zoeller: This question can be addressed within the context of section 2.3.1 of the document. These identified strengths are as follows, and

• ***Allows for a high-order neuroendocrine assessment of male reproductive and thyroid function due to the use of an intact endocrine system (i.e., HPG and HPT axes).***
The *in vivo* nature of this assay means that the interactions of hormone signaling within the HPG and HPT axes can be captured, and this is a genuine strength of the assay. If this is what is meant by "high-order neuroendocrine", then I agree. However, the role of the hypothalamus in mediating effects of chemical treatment on the endpoints captured in this assay cannot be ascertained. Thus, the term "neuroendocrine" is overstated at best.

• ***Advances scientific understanding through its MOA and, perhaps, mechanistic approach (i.e., measurement of serum concentrations of reproductive steroids, gonadotropins and thyroid hormones).***
This is an overstatement at best. It should be clear that the 15 day adult male assay cannot be considered a "mechanistic" approach to understanding new information about basic endocrinology. At best, this assay can identify a broad range of chemicals that interfere with androgen and thyroid endocrine system.

• ***Provides MOA data (e.g., differentiates between receptor and nonreceptor-mediated effects) that can be used to tailor the design of more definitive Tier-2 tests to focus on selective endpoints to accurately identify potential hazards, define dose responses, and determine the level of risk of potential endocrine disruptors.***

The 15-day intact adult male assay cannot differentiate between "receptor and nonreceptor-mediated effects" and it doesn't need to. In fact, which receptor is being addressed in this statement" the androgen, estrogen or thyroid hormone receptor? The FSH, TSH or LH receptor? Unfortunately, this statement is so naïve that it undermines the credibility of this section. The goal of the 15-day intact adult male assay is not to determine the mechanism of action, but to recruit information about the ability of a chemical to act as an EDC on the androgen or thyroid system. The best that it can hope to do is to help to reduce the number of false-negatives in the Tier-1 screen.

• ***Allows for the maximum tolerated dose (MTD) to be readily defined since mature animals are less susceptible to marked changes in growth and less susceptible to nonspecific alterations in endpoints secondary to bodyweight changes.***

This statement, taken literally, states that a strength of this assay is that adult animals are less sensitive to the toxic effects of chemicals than less mature animals. Is this a true strength? The document does not defend such a statement in any manner, which would seem to be required for such a statement.

• ***Flexible for modifying or adding apical, histological and hormonal endpoints in the context of a single assay to detect other potential endocrine-related effects as future application may dictate.***

This is a potential strength, but the methods by which an additional endpoint would be validated is not clear. Moreover, the methods by which the endpoints described in the current version of the assay are not well supported in this document or in the supporting publications. From an endocrinological perspective, important questions about the sensitivity of these endpoints to specific disruptions in endocrine action are unanswered.

• ***Complies with the basic principles of good laboratory animal practice (i.e., three R's - Reduce, Refine, and Replace), specifically through the effective use of a minimal number of animals.***

This is a clear strength of the assay. It is clear that a single in vivo assay can test the ability of unknown chemicals to interfere with endocrine action at a number of levels at once. This could not be accomplished by a single – or even multiple – in vitro assays.

• Complies with the expected simplicity and rapidity of a screen prescribed by the EDSTAC since the in-life portion of the assay is readily applied and minimal in duration.

In principle, this assay should provide rapid information about the ability of a chemical to interfere with endocrine action. The degree to which this is a strength is related to the ability of the EPA to sharpen this assay, its justification, commission and interpretation.

Assay Limitations The limitation of the assay is that the background information does not provide essential information required to interpret the results. For example, to what extent must TSH be elevated to produce histological changes in the thyroid gland? To what extent must TSH be elevated to produce thyroid tumors in the SD rat? How can total T_4 be reduced without changing serum TSH? Why do some chemicals cause a reduction in total serum T_4 without a concomitant change in serum TSH, yet other chemical produce a change in serum TSH?

2.7 Comments on the Impacts of the Choice of (Test Substances, Analytical Methods, and Statistical Methods in Terms of Demonstrating the Performance of the Assay)

Reviewer comments are organized into the corresponding sub-sections below.

2.7.1 Test Substances

George Daston: Only two test substances were evaluated in the assay, Phenobarbital and linuron. These represent a good start, as they test the ability of the protocol to detect an agent that acts indirectly on the thyroid, and a weak anti-androgen. Clearly, many more compounds that act by these and the other modes of action the test is designed to detect will need to be evaluated. However, these two were appropriate as a first step in evaluating assay reliability and transferability.

Richard Dickerson: The choice of linuron and phenobarbital were appropriate for several reasons. First, they are well-characterized EDCs with known mechanisms of

action that target two of endocrine systems of interest. Second, that is an extensive data base on these compounds to which the validation results can be compared. However, one of the stated strengths of the adult male assay is that it can detect effects on the estrogen hormone system as well as the androgen and thyroid hormone systems (page 3, Table 2). Addition of one of the weak estrogens listed in Table 4 (page11) as a test substance would increase the validity of the system. As it stands, the ISP demonstrates inter-laboratory concordance in the identification of weak or partial androgens and thyroid hormone excretion enhancers.

Kevin Gaido: The tests substances a weak androgen receptor antagonist and a compound that targets thyroid function were appropriate.

Richard Sharpe: The pre-validation studies have used compounds with a wide range of hormonal or other activities and these have provided a robust evaluation of the effectiveness of the assay and of its sensitivity and discriminatory powers. Much of this data has been obtained only in single laboratories and with considerable experience of running this assay. This may not provide an accurate guide as to how usable the assay will be when let loose in the "real world". However, these studies have served their undoubted purpose. In terms of the inter-laboratory validation exercise, I endorse the selection of linuron and phenobarbital as test compounds, as neither has profound endocrine disruptor activity comparable to a chemical such as flutamide, for example. Phenobarbital has quite major biological effects and has effects on bodyweight and the thyroid axis but was not expected to impact on the reproductive axis. It therefore provided a good choice via which to see how discriminating the assay could be in picking up thyroid changes whilst showing no effect on the reproductive axis. This was largely achieved. Similarly, selection of linuron was a good choice because its use in the assay in at least three different laboratories in the pre-validation exercise had shown reproducible effects on androgen target organs and on hormone levels but only at high doses and only in a rather selective way; again, it did not have profound activity such as a compound like flutamide. Its inclusion in the inter-laboratory validation exercise was therefore a good choice as it has the sort of activity that would make it a good candidate for becoming a "false negative". The fact that all three laboratories provided statistically significant

evidence for its 'anti-androgenicity' (based on effects on epididymal weight at one or more doses) is therefore reassuring, though I mention elsewhere my concerns about the use of epididymal weight as a definitive measure of anti-androgenicity.

Tom Zoeller: There were a large number of chemicals that could have been used, but this does not seem to be an essential issue. It is not clear why Phenobarbital and Linuron were used because both activate liver enzymes that likely cause a reduction in serum thyroid hormone, but these chemicals are as suitable as other chemicals that interfere with androgen or thyroid action. Interestingly, these chemicals illustrate a serious weakness both in the commission of the studies and in the interpretation of results. Specifically, the conclusions are based on the expected findings and not on a logical framework that was established *a priori*. This issue will be addressed elsewhere.

2.7.2 Analytical Methods

George Daston: The analytical methods were appropriate. The variability around most of the measurements was acceptable, with dihydrotestosterone and prolactin being possible exceptions. Since these were of limited use in the interpretation of the modes of action being evaluated, this did not affect the interpretability of the test results. However, if the assay is to be used for its broadest possible applications, the variability of the assays for these hormones will need to be improved.

Richard Dickerson: Most appropriate for the measurement of the hormones of interest. Other methods, such as LC-MS, may be more sensitive but are very limited in terms of sample throughput, require expensive equipment, and are not appropriate for a screening assay.

Kevin Gaido: The analytical and statistical methods appear appropriate.

Richard Sharpe: Essentially two analytical methods are used as part of the test, hormone assays and selective evaluation of organ histopathology (testes, epididymides, thyroid). If the test is to be practicable and applicable in laboratories around the world, it

demands consistency in terms of assay kits used as there is enough variation anyway in hormone measurements between laboratories when using the same assay kit. Standardization of the kits used and consequently of the method used is therefore an important step towards uniformity as well as minimizing inter-laboratory variation. However, such variation is commonplace and likely to be considerable when, and if, the assay is put into widespread use by laboratories that have little experience with running hormone assays. For this reason, organ histopathology will continue to be an important component of the test as it may provide confirmation of a target organ effect for a compound with relatively limited activity. It is perhaps not reassuring that no histopathology was picked up for linuron in any of the laboratories except for one laboratory reporting very minor testicular changes. As all of the laboratories involved are experienced in organ histopathology, it suggests that this assay will only be useable in laboratories with resident histopathology expertise and is likely to be insensitive on its own.

I am not certain of the relevance and importance of FSH measurements in the current assay in relation to its overall purpose. FSH levels will only normally increase significantly when there is quite severe impairment of spermatogenesis such that testicular weight is decreased and secretion of inhibin-B is reduced. Therefore there is no obvious benefit of measuring FSH in this situation when the information will already be provided by a more easily measurable endpoint ie. testis weight. In the inter-laboratory validation exercise, linuron exposure mildly elevated FSH levels whereas phenobarbital mildly decreased FSH levels and in neither case was it obvious why this should have occurred due to an anti-androgenic mechanism or due to any effects on the testis itself, which either did not occur (phenobarbital) or were trivial in nature (linuron). Estradiol is a negative regulator of FSH, but levels were significantly elevated in both linuron and phenobarbital-exposed animals, which thus provides no consistent explanation for the altered FSH levels. This also draws into focus why estradiol levels should be increased (there is no obvious explanation for either treatment) and what is the precise importance of estradiol measurements in the current assay? I am not sure that we yet fully understand the roles of estradiol in the male and this may make it difficult to interpret changes in estradiol levels, as for example in the inter-laboratory comparison, and what this may

mean in terms of the endocrine disrupting potential of a test compound. Additionally, blood levels of estradiol are very low in an adult male rat, are not easy to measure by assay (the assay used presently has a high coefficient of variation) and is likely to prove one of the more problematical measurements once the assay is adopted by a wider number of laboratories, especially those with little experience in running hormone assays.

I am unconvinced of the need for measurement of DHT in the present assay. The only obvious benefit of its inclusion is that it may help to identify the MOA for compounds that act as 5α reductase inhibitors. However, as such compounds should also be picked up by their effects on weight of androgen target organs, inclusion of this particular hormone assay in this Tier-1 screen is probably an unnecessary complication unless experience subsequently proves that 5α reductase inhibition is a common effect of compounds under investigation.

Tom Zoeller: EPA made a serious mistake in not setting very precise performance standards for specific methods, especially the RIA. This is the most important reason that the data appear to indicate that hormone levels are variable from one lab to another. Considering that RIAs have been in existence for nearly 50 years, it is remarkable that EPA set up such an inter-laboratory test with no apparent thought with the ways other entities (e.g., CDC) have design to allow quantitative comparisons of results between various independent laboratories. These issues are described below.

1. Radioimmunoassay (RIA). Minimally, EPA must establish performance standards for the RIA that a laboratory uses in the commission of this (and all other in vivo) assay. These performance standards should include:

i. Intra-assay variation. Each of the commercial kits reported by the laboratories in this study reported an intra-assay variation far below that reported by the laboratory using the kit. What reason could there be for EPA to accept data that do not meet the performance standard of the kit? Thus, EPA should specify that the end user of the kit is at least using the kit properly as defined in part by their reported intra-assay variation. In fact, the reported difference in performance of the assay may be related to the difference in the way in which the intra-assay

variation was defined. Companies using these kits should use the same method to define intra-assay variation as that described in the kit. If they do not, they are not actually measuring the intra-assay variation. Related to this, in the individual description of the work, one company specified that all samples from a single experiment were run within a single assay. Another company did not specify and the third reported that when all samples could not be run in a single assay, the samples were distributed across groups in different assays. This is simply unacceptable and EPA should require that all samples be run in a single assay.

ii. *Inter-assay variation*. EPA should require that companies measure inter-assay variation as it is described in the kit they are using, and that this measure falls within the limits reported by the kit manufacturer.

iii. Validation. EPA should either require that companies use homologous assay kits (i.e., designed and validated for rat), or require that companies validate the use of heterologous assays in their own lab. For example, we have demonstrated in my lab that the kit designed to measure total T_4 in humans is not valid for rat serum. Thus, these kits do not always perform well for other species and this is a minimal requirement that EPA should have anticipated. In addition, EPA should specify the validation method. This would likely include a linearity check using dilutions of rat serum in addition to "spiking" rat serum with the hormone of interest and evaluating "recovery".

iv. *Inter-laboratory consistency.* Consistency across laboratories could be better standardized if EPA developed and distributed the QC standards to which all commercial kits would have to be cross-calibrated. This is an issue that should easily have been anticipated. The Centers for Disease Control and Prevention will have a system that that the EPA could model on a smaller scale. They likely have a set of performance standards for contract (clinical) laboratories that will be very helpful.

2.7.3 Statistical Methods in Terms of Demonstrating the Performance of the Assay

George Daston: The statistical methods appeared to me to be appropriate in comparing across laboratories. Although there were important differences in interlaboratory results, the overall interpretation of study results was consistent across laboratories, a conclusion supported by the analyses.

Richard Dickerson: Not my area of expertise. I consult with a statistician to select appropriate methodology but the methods listed are the ones recommended to me for similar studies.

Kevin Gaido: The analytical and statistical methods appear appropriate.

Richard Sharpe: As far as I am able to judge, the statistical methods used for analysis of the significance of effects, for analysis of trends and for comparison of variability in methodology between laboratories is appropriate. However, I am not sure that any statistical package can truly evaluate the performance of the assay as this has to integrate all of the organ and hormonal data in a way that allows objective decision making and classification and I am not certain that this is possible. Instead, I feel that such decision making will be based not on appropriate individual statistical tests but analysis of the data by experts who have experience of the test and with results and variability in responses that it shows for different chemicals.

Tom Zoeller: I was honestly surprised to see that the statistical methods used to characterize the performance of the assay across laboratories were almost completely designed post-hoc. Perhaps the degree of variation was higher than anyone expected. Certainly, variation in hormone levels among controls could be reduced considerably by the methods described above. However, variation in treatment effects could also be from differences in feed. While the same feed was used for this 15-day study, is it possible that the animals were not on the same feed prior to this? Thus, if, as the EPA speculate, phytoestrogens in the feed would not impact adult males in 15 days, it may also follow that the feed the animals were on would affect the results when the animals were only

taken off that feed for 15 days. The point here is that so many mistakes were made in standardizing the assay across laboratories that it is doubtful that finding just the right statistical analysis will help in this particular case.

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

George Daston: My overall conclusion is that results to date are promising but not perfect. On the promising side, all three labs were able to identify the signals that would categorize linuron as an anti-androgen and Phenobarbital as affecting the thyroid. However, the concordance of results across labs for all endpoints was not perfect. In an assay with as many measurements as this one, some variability is bound to happen. It seems, however, that the results for Phenobarbital in two of the three labs was sufficiently different from expected that it is likely that, were it to be an unknown compound, it might have been erroneously classified as a male reproductive toxicant.

Given that the study evaluated only two compounds, it is difficult to make any definitive conclusions about repeatability and reproducibility. My own conclusion is that, combining the results of this study with the results described in the prevalidation section of the report, it appears that the study has reasonable transferability and reproducibility. In order to be more definitive, it will be necessary to test more chemicals. There should be more than one chemical with the same mode of action as the two already tested, as well as sets of chemicals that evaluate other modes of action. There need to be a reasonable number of non-endocrine toxicants tested at a maximally tolerated dose level to distinguish non-specific toxic responses from those that are indicative of a mode of action.

Richard Dickerson:

a. Variability in measuring hormone levels between laboratories. This is a concern particularly with the peptide hormones FSH, LH, TSH and prolactin. It is understood that these hormones are more labile than the steroid hormones and thyroid hormones. However, T is a major hormone of interest and the variability of its assay was among the highest.

b. Reproducibility in measuring hormone levels between laboratories. There were differences between laboratories in the absolute values measured for many of the hormone measurements. These differences disappeared when relative values were used and the linear trend was generally reproducible.

Potential sources of error in hormone analysis include:

A. Method of blood collection: trunk blood vs. cardiac puncture. Collection of blood following decapitation may contaminate the sample with tissue proteases that can degrade peptide hormones. Cardiac puncture or venipuncture require more skill but are less likely to contaminate blood samples with hair or tissue.

B. Blood handling and storage: Peptide hormones are much more sensitive than steroid hormones to degradation by intrinsic proteases. Although several studies suggest the sample handling procedures should ensure hormone integrity for at least 72 hours, the fact that prolactin hormone concentrations from samples collected by Charles River but analyzed by RTI were considerably lower than those determined by RTI or WIL suggest a sample handling or storage problem.

C. Method of euthanasia: Were the rats brought to the necropsy room individually or in groups? It is well-documented that the decapitation of the first rat causes stress effects that alter hormone levels in other rats present in the room. Prolactin and corticosterone appear most sensitive to stress. The AVMA guidelines on euthanasia suggest that the sounds and odors associated with euthanasia can result in the release of stress hormones in other animals in that room. Although glucocorticoids are not a focus of this assay, prolactin is an endpoint.

Kevin Gaido: There were inconsistencies in hormonal measurements between laboratories and relative to historical controls. In addition, the effects of linuron on various endpoints varied between laboratories. However based on a weight of evidence approach, each laboratory correctly identified linuron as an androgen receptor antagonist and phenobarbital as a compound that interferes with thyroid function. Thus, while the variability associated with this screening assay is a weakness, the inclusion of multiple endpoints strengthens this assay and increases its reliability.

Richard Sharpe: From the published studies and reports of the pre-validation investigations, the reproducibility of the assay appears to be very good. Nevertheless this has only been investigated for one or two compounds and it would be a major surprise if the effects of a potent anti-androgen such as flutamide, for example, were not picked up robustly in the assay when used by different laboratories. It is beyond dispute that the assay works effectively in the various test situations in which it has been applied, but the challenge remaining is to decide how it will be operated in practice by different laboratories in the future and what will be the false negative and false positive rates. Will "linuron-like" compounds always be positively identified? In addressing this question for linuron itself, it is reassuring that its effects have been picked up in the assay by different laboratories both in the pre-validation studies and in the inter-laboratory validation exercise. This imparts a degree of confidence that similar compounds, for which the activity is completely unknown, will be picked up during future use of this assay by different laboratories. Furthermore, the studies with phenobarbital reinforce the view that the assay is robust, as it was negative in all of the laboratories for reproductive organ endpoints whilst showing positive activity for thyroid gland weight and thyroid hormone changes. Nevertheless, it is somewhat disconcerting that all three laboratories identified significant effects on blood levels of FSH and estradiol after phenobarbital exposure, the changes with FSH even being observed at the lowest treatment dose. The most rational explanation for this observation is that it is the elevation in estradiol levels that is responsible for the suppression of FSH, but the question then is whether phenobarbital would be classified as an endocrine disruptor of the reproductive system based on these

observations? I am not entirely sure what the purpose of inclusion of FSH and estradiol measurements is in the current assay, and this is a point that is discussed elsewhere.

The changes in thyroid gland weight and thyroid hormone levels appear particularly robust in the pre-validation studies and especially in the inter-laboratory validation exercise. I do not feel sufficiently expert to comment in any informed way on the value of the assay as applied to the thyroid axis, but I am disappointed to note that only one of the laboratories was able to confirm the thyroid hormone changes with appropriate histopathological changes; one of the other laboratories also noted minor changes but only at the very highest dose. This questions how useful target organ histopathology is going to prove in this assay, when it cannot match up with the hormone data which is changing in such a reproducible fashion.

Tom Zoeller: I do not believe that the variability observed in the assay is inherent in the biological and chemical test methods. Unfortunately, this exercise has failed to answer this question.

2.9 Additional Comments and Materials Submitted

George Daston: Literature Cited

Cook, JC, Kaplan, AM, Davis, LG, O'Connor, JC 1997. Development of a tier 1 screening battery for detecting endocrine-active compounds (EACs), Regul. Toxicol. Pharmacol. 26: 60-68.

O'Connor, JC, Cook, JC, Craven, SC, Van Pelt, CS, Obourn, JD 1996. An in vivo battery for identifying endocrine modulators that are estrogenic or dopamine regulators. Fundam. Appl. Toxicol. 33: 182-195.

Tom Zoeller: REFERENCES

1. **vom Saal FS, Richter CA, Mao J, Welshons WV 2**005 Commercial animal feed: variability in estrogenic activity and effects on body weight in mice. Birth Defects Res A Clin Mol Teratol 73:474-475

2. **Vom Saal FS, Richter CA, Ruhlen RR, Nagel SC, Timms BG, Welshons WV** 2005 The importance of appropriate controls, animal feed, and animal models in interpreting results from low-dose studies of bisphenol A. Birth Defects Res A Clin Mol Teratol 73:140-145

3. **Blake NG, Johnson MR, Eckland DJ, Foster OJ, Lightman SL** 1992 Effect of food deprivation and altered thyroid status on the hypothalamic-pituitary-thyroid axis in the rat. J Endocrinol 133:183-188

4. **Blake NG, Eckland DJ, Foster OJ, Lightman SL** 1991 Inhibition of hypothalamic thyrotropin-releasing hormone messenger ribonucleic acid during food deprivation. Endocrinology 129:2714-2718

5. **Coppola A, Hughes J, Esposito E, Schiavo L, Meli R, Diano S** 2005 Suppression of hypothalamic deiodinase type II activity blunts TRH mRNA decline during fasting. FEBS Lett 579:4654-4658

6. **Diano S, Naftolin F, Goglia F, Horvath TL** 1998 Fasting-induced increase in type II iodothyronine deiodinase activity and messenger ribonucleic acid levels is not reversed by thyroxine in the rat hypothalamus. Endocrinology 139:2879-2884

Table 2-1. Tom Zoeller: Effect of Linuron on Selected Hormones in the three studies.					
	Site	**Control**	**50mg/kg**	**100 mg/kg**	**150 mg/kg**
Testosterone (ng/mL)	WIL	6.137 ±0.95	6.317 ± 1.267	6.225 ± 1.267	4.102 ± 1.312
	RTI	3.37 ± 0.75	4.07 ± 0.42	2.56 ± 0.42	2.49 ± 0.43
	Charles-River	9.93 ± 1.873	4.83 ± 0.960	3.98 ± 1.065	3.28 ± 0.641
LH (ng/mL)	WIL	0.693 ± 0.098	0.880 ± 0.098	0.667 ± 0.098	0.629 ± 0.102
	RTI	1.28 ± 0.08	1.33 ± 0.07	1.31 ± 0.08	1.41 ± 0.08
	Charles-River	2.18 ± 0.127	1.75 ± 0.095	1.78 ± 0.137	1.86 ± 0.179
T_4 (µg/dL)	WIL	4.973 ± 0.217	2.687 ± 0.229	1.287 ± 0.205	0.521 ± 0.064
	RTI	5.55 ± 0.16	3.87 ± 0.16	2.61 ± 0.16	1.69 ± 0.17
	Charles-River	4.73 ± 0.151	3.10 ± 0.151	1.82 ± 0.151	1.54 ± 0.151
T_3 (ng/dL)	WIL	79.993 ± 2.859	78.627 ± 2.859	69.873 ± 2.859	66.943 ± 2.96
	RTI	87.53 ± 2.92	85.66 ± 2.92	76.90 ± 2.92	77.98 ± 3.02
	Charles-River	81.65 ± 2.593	64.85 ± 2.593	65.41 ± 2.593	56.15 ± 2.684
TSH (ng/mL)	WIL	15.393 ± 1.697	13.947 ± 1.357	12.247 ± 1.357	11.386 ± 1.41
	RTI	18.53 ± 3.18	18.51 ± 1.80	11.90 ± 0.88	13.85 ± 1.53
	Charles-River	13.10 ± 1.681	23.35 ± 2.877	25.74 ± 2.024	29.73 ± 2.269

Table 2-2. Tom Zoeller: Effect of Phenobarbital on Selected Hormones in the three studies.					
	Site	**Control**	**25 mg/kg**	**50 mg/kg**	**100 mg/kg**
Testosterone (ng/mL)	WIL	6.137 ±0.95	4.475 ± 0.674	3.501 ± 0.674	2.870 ± 0.698
	RTI	3.37 ± 0.75	2.62 ± 0.52	2.64 ± 0.52	2.66 ± 0.56
	Charles-River	9.93 ± 1.873	4.83 ± 0.960	3.98 ± 1.065	3.28 ± 0.641
LH (ng/mL)	WIL	0.693 ± 0.098	0.647 ± 0.098	0.560 ± 0.098	0.779 ± 0.102
	RTI	1.28 ± 0.08	1.28 ± 0.07	1.24 ± 0.08	1.18 ± 0.08
	Charles-River	2.18 ± 0.127	1.81 ± 0.098	1.44 ± 0.054	1.56 ± 0.068
T_4 (µg/dL)	WIL	4.973 ± 0.217	4.133 ± 0.293	2.893 ± 0.148	1.707 ± 0.204
	RTI	5.55 ± 0.16	4.66 ± 0.16	4.32 ± 0.16	2.97 ± 0.17
	Charles-River	4.73 ± 0.151	3.75 ± 0.151	3.64 ± 0.151	2.62 ± 0.156
T_3 (ng/dL)	WIL	79.993 ± 2.859	72.040 ± 2.859	62.773 ± 2.859	57.086 ± 2.96
	RTI	87.53 ± 2.92	78.24 ± 2.92	73.28 ± 2.92	58.14 ± 3.14
	Charles-River	81.65 ± 2.593	64.85 ± 2.593	65.41 ± 2.593	56.15 ± 2.684
TSH (ng/mL)	WIL	15.393 ± 1.697	24.013 ± 2.816	25.833 ± 2.816	31.857 ± 2.92
	RTI	18.53 ± 3.18	21.69 ± 1.79	26.6 ± 2.42	26.38 ± 3.50
	Charles-River	13.10 ± 1.681	23.35 ± 2.877	25.74 ± 2.024	29.73 ± 2.269

3.0 PEER REVIEW COMMENTS ORGANIZED BY REVIEWER

Peer review comments received for the 15-day intact adult male rat 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.

3.1 George Daston Review Comments

Comments on Integrated Summary Report for Validation of 15-Day Intact Adult Male Rat Assay as a Potential Screen in the Endocrine Disrupter Screening Program Tier-1 Battery

George Daston

September 21, 2007

1. Please comment on the clarity of the stated purpose of the assay:
In order to provide the reader with an understanding of the purpose of the assay, it is necessary first to provide the context in which it will be used. The summary report does a good job of explaining the legislative mandate for endocrine screening, the tiered approach that EPA has decided to take, and the aspects of the screening tier that are germane to the development of the adult male assay. The only niggling issue that I had with the presentation of regulatory context is the statement that the legislative mandate is part of the Federal Food Drug and Cosmetic Act (FFDCA). It has been explained to me that this is technically correct; however, most of us consider the endocrine disrupter screening program to be a mandate of the Food Quality Protection Act. While I now understand that the FQPA made modifications to both FIFRA and FFDCA, this point escaped me when I first read the report. This confusion is compounded by language on line 13, p. 1 "Subsequent to passage of the Act in 1996", with "Act" referring to FFDCA, a law that was passed more then 90 years previously. It would be an easy fix to add a phrase indicating that the regulatory statute is FFDCA *as modified by the Food Quality Protection Act of 1996*.

The report's interpretation of validation of alternative tests under ICCVAM has a few inaccuracies that should be corrected. These have to do with the interpretation that the validation

process was intended specifically for in vitro replacements of in vivo assays (p. 4, line 3, line 45; p. 49, lines 8-9). This is not the intention of the ICCVAM criteria. The criteria are intended to assess whether any assay -- in vivo, in vitro, in silico – is sufficiently robust to serve as an alternative to an existing test method that has regulatory acceptance. ICCVAM has reviewed and accepted in vivo methods as alternatives, including the up-down method for acute toxicity and the local lymph node assay for contact allergy. I don't agree that the ICCVAM criteria represents a "fundamental problem confronting the EPA" as is stated on lines 3-4, p. 4. The major difference between the validation for the endocrine assays and that of other assays is the absence of a gold-standard assay with a large database against which to compare results. This latter problem is the one that the report tries to grapple with, and I agree that it is a legitimate issue. For the sake of clarity in the organization of the report, it would be much preferable to scrap the spurious argument that the validation process is designed for in vitro tests and to acknowledge that because the endocrine screening assays aren't replacing a specific test method some flexibility will be required in how the validity of the new test methods are interpreted.

Given that the purpose of the assay is to identify specific modes of endocrine toxicity, I believe that the correct approach is to validate the performance of the assay using a set of compounds for which the modes of action have been generally agreed upon through the development of a large data set in the literature. The report tries to do this, but it would be much easier to follow if this were presented as the context of providing a standard for validation. For example, one would classify linuron as having anti-androgenic activity or Phenobarbital as having thyrotoxic activity based on a critical review of the literature. This review includes the verification that this activity has adverse consequences on the male reproductive system or thyroid, respectively, in a toxicity study that conforms to regulatory guidelines. This approach would satisfy the validation criteria and provide a basis for making calculations of assay performance (e.g., concordance, sensitivity, specificity, etc.).

One of challenges for the report's authors is to clearly present information on how to interpret a test as complicated as this one. The report could be better in this respect. I would especially like to see a section that provides criteria for interpretation. There are perhaps a dozen modes of action that this assay was designed to detect: thyroid disturbances, androgen receptor agonists and antagonists, estrogen receptor agonists and antagonists, progesterone receptor agonists and

antagonists, inhibitors or enhancers of steroid synthesis, dopamine agonists and antagonists (assessed via prolactin modulation) and other modes that perturb the pituitary response within the hypothalamic-pituitary-gonadal axis or the hypothalamic-pituitary-thyroidal axis. Ideally what I would like to see is a short description of which assay endpoints would be changed in order to categorize something as an androgen antagonist, steroid synthesis inhibitor, etc. This could be done in the text, or as a flow chart. If it is necessary to include information from other tier 1 screening assay, that's fine, as it appears from Table 3 that this test would be performed as part of a battery. I realize that there will be a need to modify these interpretations as more data become available for this test method. However, the endpoints in the assay were selected based on solid mechanistic understanding that by measuring them it would be possible both to detect certain modes of action and rule out others. The developers of these assays have been stating such interpretations since first publishing on these tests in the '90s (O'Connor et al., 1996; Cook et al., 1997). It would be useful to have short summaries here. The appropriate place would be at the end of section 3.

The intact male assay has been used extensively by a number of industry labs. The data from these labs is summarized nicely in the Prevalidation section of the report. However, the industry groups appear to be using the assay for a broader range of modes of action than EPA evaluated in its validation study. Table 4 lists such modes of action as progesterone receptor agonism/antagonism, dopamine receptor agonism/antagonism, that were beyond the scope of the validation program. It wasn't clear to me in reading the report whether this was simply due to the limited scope of the initial validation, or if EPA intends to scale back the purpose of the study. I would like to see this point addressed specifically, in section 2.1.

2. *Please comment on the clarity, comprehensiveness and consistency of the data interpretation with the stated purpose of the assay*:

The primary purpose of the study, as described in section 4.1, was to evaluate the reliability and transferability of the newly developed standard protocol, and to a lesser extent to continue to assess assay relevance. Given the primary purpose of the study, I believe that the right data are emphasized in section 5 of the report. The authors of the report stayed focused on the goals of transferability, reliability, and adherence to protocol.

There are comparisons to historical control data, particularly for body and organ weight, that might be interpreted differently if additional historical control data were considered (Table 9 and the accompanying text). The historical control data appear to be limited to 28 studies using a similar study design and compiled by O'Connor et al in 2002. Many of the studies in O'Connor's paper were several years old at the time, and are now more than 10 years old. There is a constant, subtle drift in body and organ weights over time such that the older data may not be as relevant. Furthermore, it isn't possible to know whether subtle differences in housing conditions or husbandry in the various labs produces variability in relative organ weights. Therefore, it would be useful to include historical control data from each of the three labs for this species and strain of rat. It is likely that they have data for control body weights for SD rats from 10-12 weeks of age because this is within the age range of animals used for subchronic and reproductive toxicity studies. The age-range is a little young for organ weight data from 91-day subchronic studies, but relative organ weight (organ/body weight) may be informative.

3. *Please comment on the biological and toxicological relevance of the assay as related to its stated purpose*.

I believe that the biological and toxicological relevance of the assay is well described in section 3.1. I believe that this assay, as well as the pubertal male and pubertal female assays being evaluated, has the potential to provide the most reliable and comprehensive information for the weight-of-evidence determination described in section 1.4. The use of an intact animal model provides the opportunity to assess multiple endocrine processes, both alone and in integration with the hypothalamic-pituitary axes that control thyroid and gonadal function. The ability to measure multiple modes of action in a single assay provides the opportunity to obtain a lot of information from a relatively small number of animals, vs. running separate tests for each mode of action. The intactness of the hypothalamic-pituitary-gonadal and hypothalamic-pituitary-thyroidal axes makes the model biologically relevant, as these axes act in concert in the organism that we wish to model for the purposes of hazard and risk assessment, the human. The model is toxicologically relevant because the responses in an intact system, which also has homeostatic mechanisms, is likely to be much more concordant with the results of more definitive toxicity tests.

4. Please 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

Most of the information regarding the purpose of the assay is in the Introduction. I found that this section had some of the same clarity problems as the report. Specifically, it is not clear whether the purpose of this standardized protocol is to conduct an assay that is capable of detecting all of the modes of action listed in the first paragraph (p. 4 of the protocol), just estrogen, androgen, or thyroid-related modes, as is implied in the second paragraph, or the list of modes described in the third paragraph: AR agonists and antagonists, steroid biosynthesis inhibitors, gonadotropin and thyroid modulators. It would be easier for the lab to understand the nature of the work if two of these three were eliminated from the introduction to the protocol.

b. conduct the assay

The right information is present in the protocol, as evidenced by the fact that none of the three labs had any serious deviations from protocol.

c. observe and measure prescribed endpoints

The experimental design is very detailed and specific. The procedures seem clear and interpretable.

d. compile and prepare data for statistical analysis

The procedures for data compilation and statistical analysis are clear.

e. report results

The protocol is very clear about the data that should be summarized, even down to the point of prescribing which data should be in tables, which in figures, and how the figures should be drawn. This level of control over data presentation is more than what I am accustomed to seeing. I believe that it is useful to have this level of control when making head-to-head comparisons of interlaboratory results. However, it may be prudent to remove it should the assay become routine.

The section on interpretation of effects (pp. 15- 18 of the protocol) was surprising to find in the protocol. Given that the primary purpose of the study was to determine reliability and transferability, it would seem to me that this kind of information is more relevant to the study sponsors than the participating labs. I question that it should be in the protocol.

5. *Please comment on the strengths and limitations of the assay*

The strengths of the assay are nicely laid out in section 2.3 of the report. As noted in my response to question 3, I consider the fact that this is an intact system is a major strength. It should be possible, with a limited number of measurements, to obtain information on a number of modes of endocrine action. The level of information may actually be more than what one can obtain from many definitive toxicity tests. This assay (or its alternates, the pubertal male or pubertal female assay) will provide the greatest weight in weight-of-evidence schemes that will be applied to the tier 1 battery.

I agree with the limitations and challenges given in section 2.3.2. I would add one or two more. First, it is not clear yet whether this assay will have the sensitivity that other assays have, largely because it is an intact model. The fact that there are homeostatic mechanisms in place will tend to blunt – not overcome, but blunt – some of the responses that are being examined. One of the potential strengths of this assay is that some of the potential for homeostasis can be unmasked through the measurement of hormone levels.

The hormone measurements are among the more important aspects of this protocol, especially if one of the goals of the assay is to obtain a mode of action fingerprint. I take the results of the interlab comparisons of hormone levels to be promising, except for two of the hormones which are probably the least important contributors to the resolving power of the assay, at least for the modes of action it will be applied to in the EPA tier 1 battery.

I found the apparent lack of specificity of T3 and T4 as indicators of thyroid toxicity to be troubling. It will be necessary to develop much more data on negative compounds (i.e., non-endocrine disrupters) in this assay to ensure that changes in thyroid hormone levels can be appropriately interpreted.

6. *Please provide comments on the impacts of the choice of:*

a. test substances

Only two test substances were evaluated in the assay, Phenobarbital and linuron. These represent a good start, as they test the ability of the protocol to detect an agent that acts indirectly on the thyroid, and a weak anti-androgen. Clearly, many more compounds that

act by these and the other modes of action the test is designed to detect will need to be evaluated. However, these two were appropriate as a first step in evaluating assay reliability and transferability.

b. analytical methods

The analytical methods were appropriate. The variability around most of the measurements was acceptable, with dihydrotestosterone and prolactin being possible exceptions. Since these were of limited use in the interpretation of the modes of action being evaluated, this did not affect the interpretability of the test results. However, if the assay is to be used for its broadest possible applications, the variability of the assays for these hormones will need to be improved.

c. statistical methods in terms of demonstrating the performance of the assay

The statistical methods appeared to me to be appropriate in comparing across laboratories. Although there were important differences in interlaboratory results, the overall interpretation of study results was consistent across laboratories, a conclusion supported by the analyses.

7. *Please provide comments on the repeatability and reproducibility of the results obtained with the assay, considering the variability inherent in the biological and chemical test methods.*

My overall conclusion is that results to date are promising but not perfect. On the promising side, all three labs were able to identify the signals that would categorize linuron as an anti-androgen and Phenobarbital as affecting the thyroid. However, the concordance of results across labs for all endpoints was not perfect. In an assay with as many measurements as this one, some variability is bound to happen. It seems, however, that the results for Phenobarbital in two of the three labs was sufficiently different from expected that it is likely that, were it to be an unknown compound, it might have been erroneously classified as a male reproductive toxicant.

Given that the study evaluated only two compounds, it is difficult to make any definitive conclusions about repeatability and reproducibility. My own conclusion is that, combining the results of this study with the results described in the prevalidation section of the report, it appears that the study has reasonable transferability and reproducibility. In order to be more definitive, it will be necessary to test more chemicals. There should be more than one chemical with the same mode of action as the two already tested, as well as sets of chemicals that evaluate other modes

of action. There need to be a reasonable number of non-endocrine toxicants tested at a maximally tolerated dose level to distinguish non-specific toxic responses from those that are indicative of a mode of action.

Literature Cited

Cook, JC, Kaplan, AM, Davis, LG, O'Connor, JC 1997. Development of a tier 1 screening battery for detecting endocrine-active compounds (EACs), Regul. Toxicol. Pharmacol. 26: 60-68.

O'Connor, JC, Cook, JC, Craven, SC, Van Pelt, CS, Obourn, JD 1996. An in vivo battery for identifying endocrine modulators that are estrogenic or dopamine regulators. Fundam. Appl. Toxicol. 33: 182-195.

3.2 Richard Dickerson Review Comments

CHARGE QUESTIONS

Each peer reviewer is asked to review the Integrated Summary Report and accompanying support materials and comment on the results of the validation process of the 15-day intact adult male rat assay, especially the interlaboratory validation exercise.

1. Please comment on the clarity of the stated purpose of the assay.
   First, the purpose of the report and assay should be stated much closer to the beginning of the report than on page 6. The assay and its validation are the focus of the report, not the history of why it is needed. I suggest placing section 1.6 (Purpose of the ISP) as 1.1 followed by the purpose of the assay (2.1) followed by the remainder of the introduction. This allows those individuals familiar with EDSTAC to focus on the purpose of the report and assay and perhaps skip the historical background. Another option might be to include an executive summary following the cover sheet that summarizes the purpose of the report, the purpose of the assay and the conclusions derived from the results of the interlaboratory

validation. If the target audience is the decision makers, putting the bottom line up front provides greater assurance that they will get the message.

Second, the purpose of the assay could be more directly and clearly stated. It is stated in passive voice rather than active, and begins with a reference to other publications. It is more effective to state "The purpose of the 15-day intact adult male rat assay is to detect compounds or mixtures that alter the HPE, HPA and HPT through the most probable MOA." The list of MOAs can follow along with the endpoints measured. A brief description of assay methodology can be included.

Third, if the third paragraph (beginning on line 8 of page 7) is to be included with the purpose of the assay, consider adding a reference to Table 4. This allows a quick comparison of assay capabilities.
A somewhat related comment is that a discussion of progesterone and RU486 was not included in section 3.1.1 Positive Test Chemicals. This should be added.

2. Please comment on the clarity, comprehensiveness and consistency of the data interpretation with the stated purpose of the assay.

   The data from each of the laboratories was presented clearly and factually. The CVs both within a laboratory and between the three test laboratories were adequately analyzed and discussed. The largest area of concern, i.e. the large variability in certain hormones, was identified and thoroughly discussed. The results of each of the assays were also correctly interpreted and sources of error identified. The sets of results from the linuron exposure and the phenobarbital exposure were presented in the same format. However, analysis of the results and interpretation were more thorough for linuron than phenobarbital.

   For the linuron exposure study, unacceptable variability in the results of the assays for prolactin (PRL), testosterone (T), dihydrotestosterone (DHT) and at the highest dose thyroxin (T4) occurred. It is also of concern that for some of these endpoints either no change was observed when historical data had reported an effect or an effect was observed when historical data suggested no change. However, there were no instances

where the direction of change was opposite to those previously reported. For many of the androgenic endpoints, one or more laboratories failed to detect an effect although previous studies had found decreases. It is interesting that the results obtained by Charles River more frequently matched the historical trend whereas the results for RTI did not despite the fact that RTI performed hormonal analyses for both.

For the phenobarbital exposure, unacceptable variability in the assays for T, DHT and PRL occurred at all dose levels. However, the thyrodogenic endpoints and the liver weight changes were significant as predicted by other studies.

3. Please comment on the biological and toxicological relevance of the assay as related to its stated purpose.

In terms of biological relevance, the assay endpoints reflect measures of the integrity of the hypothalamic-pituitary- androgen (HPA) and -thyroid (HPT) axes. These include changes in tissue weight, histology, and circulating hormone levels. Other assays relevant to the androgen axis might include rate of sperm production, sperm motility, and ability to undergo the acrosome reaction. However, the length of the cycle for sperm production greatly exceeds the 15-day period of chemical exposure used in this assay. Other measures of reproductive capacity also require much longer times of exposure than used for this assay. The endpoints used for the HPT axis are also the most appropriate for the length of the assay.

In terms of toxicologic relevance, the endpoints selected for the 15-day Adult Male Rat Assay are appropriate for several reasons. First, they reflect biologically relevant endpoints as discussed above. Second, previous studies using known androgen receptor agonists and antagonists demonstrate these endpoints are altered by exposure to methyl testosterone, vinclozolin, flutamide, p,p'-DDE and other AR agonist/antagonists. Finally, the endpoints are relevant because competent investigators, whether from industry, contract laboratories or academia are capable of measuring them in a consistent manner.

4. Please 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.

   My comments are based on the protocol appended as C since section 4.0 states it is the final, standardized protocol.

   a. comprehend the objective- Under objective the statement <enter the specific purpose of the assay> appears. It is therefore not possible to evaluate the clarity and conciseness of the objective. The section on personnel is also incomplete.

   b. conduct the assay- The instructions on how to conduct the assay are complete and clear. However, certain areas are troublesome. First, the dosing solutions are made in 0.25% methylcellulose in water for this assay but are prepared in corn oil for the pubertal male assay. What is the reason for this inconsistency? In addition, many water supplies have measurable amounts of perchlorate. Although these are usually below a level of concern for the general public, it should a mandatory analysis for this assay. If perchlorate is detected, the animals should not receive water from this source. Feed samples should be analyzed for phytoestrogens, all food for a given study should be from the same lot, and it would preferable for all laboratories to use the same food source. A feed low in phytoestrogens would be better than standard rodent chow. In terms of euthanasia, other animals should not be present in the necropsy room when an animal is euthanized or necropsied. A number of studies have demonstrated that when animals in the room when another rat is euthanized or necropsied experience significant increases in corticosterone and prolactin. If all the rats are in the room, the stress hormone levels will be markedly different between the first animal euthanized and the last. Although transporting the animals together minimizes one source of variability, it introduces another source if the animals are all in the necropsy room. The protocol does not specify whether the euthanasia chamber is to be precharged with carbon dioxide gas or if it will

be added slowly. The protocol does not specify whether pure carbon dioxide is used or if a mixture of carbon dioxide and oxygen is to be used. A specific technique should be utilized in all studies and it must conform to the most recent AVMA guidelines for euthanasia. In terms of hormone assays, what percentage of samples will be run as true duplicates? In addition, most RIA kits use I125 which has a relatively short half life. Using a fresh kit for some of the samples and an older kit for other samples can introduce variability. The protocol should specify that sufficient kits with the same lot number should be ordered so that all assays for a particular hormone are more consistent. Perhaps a standard sample could be prepared and sent to the laboratories as an additional QC standard. Last, samples shipped from one laboratory to another require detailed chain of custody documentation and if possible a data logger so that sample temperature and time can be documented. Minimal standards for transit time and temperature set.

c. observe and measure prescribed endpoints- clear and concise

d. compile and prepare data for statistical analyses- clear and concise but consider specifying statistical software.

e. report results- clear and concise.

5. Please comment on the strengths and/or limitations of the assay.

Strengths of the assay include ease of conducting the assay and measuring the endpoints, the short duration of exposure, biological relevance of endpoints and robust database.

Limitations of the assay are its inability to determine more downstream effects such as sperm production, motility and fecundity. However, assays that detect these endpoints are more appropriate for Tier-2. The intra- and interlaboratory variability in the hormone assays make it more difficult to detect subtle changes with any degree of significance.

6. Please provide comments on the impacts of the choice of:

a. test substances- The choice of linuron and phenobarbital were appropriate for several reasons. First, they are well-characterized EDCs with known mechanisms of action that target two of endocrine systems of interest. Second, that is an extensive data base on

these compounds to which the validation results can be compared. However, one of the stated strengths of the adult male assay is that it can detect effects on the estrogen hormone system as well as the androgen and thyroid hormone systems (page 3, Table 2). Addition of one of the weak estrogens listed in Table 4 (page11) as a test substance would increase the validity of the system. As it stands, the ISP demonstrates inter-laboratory concordance in the identification of weak or partial androgens and thyroid hormone excretion enhancers.

b. analytical methods- Most appropriate for the measurement of the hormones of interest. Other methods, such as LC-MS, may be more sensitive but are very limited in terms of sample throughput, require expensive equipment, and are not appropriate for a screening assay.

c. statistical methods in terms of demonstrating the performance of the assay- Not my area of expertise. I consult with a statistician to select appropriate methodology but the methods listed are the ones recommended to me for similar studies.

7. Please provide comments on repeatability and reproducibility of the results obtained with the assay, considering the variability inherent in the biological and chemical test methods.

a. Variability in measuring hormone levels between laboratories. This is a concern particularly with the peptide hormones FSH, LH, TSH and prolactin. It is understood that these hormones are more labile than the steroid hormones and thyroid hormones. However, T is a major hormone of interest and the variability of its assay was among the highest.

b. Reproducibility in measuring hormone levels between laboratories. There were differences between laboratories in the absolute values measured for many of the hormone measurements. These differences disappeared when relative values were used and the linear trend was generally reproducible.

Potential sources of error in hormone analysis include:

A. Method of blood collection: trunk blood vs. cardiac puncture. Collection of blood following decapitation may contaminate the sample with tissue proteases that can degrade peptide

hormones. Cardiac puncture or venipuncture require more skill but are less likely to contaminate blood samples with hair or tissue.

B. Blood handling and storage: Peptide hormones are much more sensitive than steroid hormones to degradation by intrinsic proteases. Although several studies suggest the sample handling procedures should ensure hormone integrity for at least 72 hours, the fact that prolactin hormone concentrations from samples collected by Charles River but analyzed by RTI were considerably lower than those determined by RTI or WIL suggest a sample handling or storage problem.

C. Method of euthanasia:
Were the rats brought to the necropsy room individually or in groups? It is well-documented that the decapitation of the first rat causes stress effects that alter hormone levels in other rats present in the room. Prolactin and corticosterone appear most sensitive to stress. The AVMA guidelines on euthanasia suggest that the sounds and odors associated with euthanasia can result in the release of stress hormones in other animals in that room. Although glucocorticoids are not a focus of this assay, prolactin is an endpoint.

3.3 Kevin Gaido Review Comments

INDEPENDENT PEER REVIEW OF THE 15-DAY INTACT ADULT MALE RAT ASSAY AS A POTENTIAL SCREEN IN THE ENDOCRINE DISRUPTOR SCREENING PROGRAM (EDSP) TIER-1 BATTERY

1. Please comment on the clarity of the stated purpose of the assay.

The stated purpose of the assay, as an alternative to the female pubertal assay to detect chemicals that interfere with androgen or thyroid function, or through the HPG axis is clearly stated. The goal is to develop a relatively quick, reliable screening assay that will be part of a comprehensive battery of tests for endocrine active chemicals.

2. Please comment on the clarity, comprehensiveness and consistency of the data interpretation with the stated purpose of the assay.

The summary statement provides a clear and comprehensive interpretation of the data. A detailed comparison of the results from each laboratory together with historical data is provided. To allow for sufficient interpretation of the results.

3. Please comment on the biological and toxicological relevance of the assay as related to its stated purpose.

   As stated above, the assay was designed to detect chemicals that interfere with androgen or thyroid function or with the HPG axis. While of little biological relevance, this assay is highly relevant for toxicological screening for endocrine active chemicals.

4. Please provide comments on the clarity and conciseness of the protocol in describing the methodology of the assay such that the laboratory can:

The protocol is clear and comprehensive. The objective is clearly stated and sufficient detail is presented to allow a laboratory with the appropriate expertise to conduct the assay and accurately analyze and report the results.

5. Please comment on the strengths and/or limitations of the assay.

Strengths of this assay include the ability to screen for multiple modes of action in an adult animal. The assay has multiple sensitive endpoints that can be used to help design more definitive Tier-2 testing. Because it is in vivo the assay allows for consideration of absorption, distribution, metabolism, and excretion. The assay is relatively rapid and has been standardized so that it can be performed in any laboratory that has the appropriate expertise and experience.

A weakness of this assay is the necessity of blood hormone measurements. These measurements are highly variable, inconsistent and subject to experimental conditions. Hormone measurements are not routinely done in toxicology studies and many laboratories will not have the appropriate expertise. The inconsistent results across laboratories with linuron suggests that this assay may not be reproducible for weak androgen receptor agonists. In addition, previously published studies indicate that this assay may not be a sensitive screen for week estrogens.

6. Please provide comments on the impacts of the choice of:

The tests substances a weak androgen receptor antagonist and a compound that targets thyroid function were appropriate. The analytical and statistical methods appear appropriate.

   a. test substances,

b. analytical methods, and
c. statistical methods in terms of demonstrating the performance of the assay.

7. Please provide comments on repeatability and reproducibility of the results obtained with the assay, considering the variability inherent in the biological and chemical test methods.

There were inconsistencies in hormonal measurements between laboratories and relative to historical controls. In addition, the effects of linuron on various endpoints varied between laboratories. However based on a weight of evidence approach, each laboratory correctly identified linuron as an androgen receptor antagonist and phenobarbital as a compound that interferes with thyroid function. Thus, while the variability associated with this screening assay is a weakness, the inclusion of multiple endpoints strengthens this assay and increases its reliability.

3.4 Richard Sharpe Review Comments

GENERAL COMMENTS

I have ordered my comments below according to the questions posed to reviewers. However, my placing of some comments is rather arbitrary as, in several instances, there is overlap or uncertainty in my mind as to which, if any, of the questions posed they address.

1. Please comment on the clarity of the stated purpose of the assay

The background information and discussion provided give a clear view of what the assay is intending to achieve and why it has (most of) its component parts. It is a Tier-1 assay and, as such, its priority is to maximize the detection of endocrine active compounds whilst minimizing false negatives. The use of multiple endpoints is designed to ensure this. Its particular strength, discussed in more detail later, is that it should sidestep issues related to hormone homeostasis, which is always likely to be the main confounder in an assay such as this which uses an intact animal with normally functioning homeostatic hormone systems.

I found the information on the purpose of the assay and its background to be clearly presented, easily understandable and to make commonsense. It should perhaps emphasize that the assay is not intended to be definitive, as this is important when considering results

from individual laboratories (for example, in the inter-laboratory comparison) in which inconsistency in results may occur, but in which the assay achieves its primary objective.

2. **Please comment on the clarity, comprehensiveness and consistency of the data interpretation with the stated purpose of the assay**

   Considerable data on this assay has been collected involving several laboratories and a large number of compounds with a wide variety of mechanisms of action (MOA). The evidence presented in reports and publications, primarily those by O'Connor *et al*, substantiate the view that this assay is fit for purpose. An important point that is made repeatedly, and which cannot be overemphasized, is that this assay intentionally uses multiple endpoints in order that it may more readily identify compounds with weak activity or with a profile of activity that does not fit within expected boundaries (for example a compound that exhibits both anti-androgenic and anti-thyroidal activity). The other purpose of the multiple endpoints is to provide preliminary information on the potential MOA, which may then guide decisions about subsequent testing in Tier-2. However, in my opinion, the main importance of the inclusion of multiple endpoints in this Tier-1 assay is to maximize the likelihood of detection of endocrine active chemicals whilst minimizing the chance of false negatives.

   The interpretation of the results obtained using this assay in the different laboratories, including the inter-laboratory validation exercise, are rational and fit with current understanding of how the various endocrine systems operate within the body. Every aspect of the data has been evaluated in terms of its robustness, its reproducibility, sensitivity of detection and consistency with other results in the same assay from the same laboratory or with results from other laboratories. There are some minor issues in relation to homeostatic changes (see my comments to question 5) and there are issues in relation to interpretation of weight changes for the epididymis, but these do not affect the overall conclusion that the assay is robust, but with some limitations. In making these comments, I base them very much on the pre-validation studies that involved extensive testing of a wide range of compounds rather than on the inter-laboratory validation exercise. If my evaluation was based on the latter alone, I would be less enthusiastic about the utility of the assay and again I discuss this further in relation to question 5 below.

Although the O'Connor studies using chemicals with well characterized anti-androgenic activity via one or more mechanisms (flutamide, ketoconazole and finasteride) are highly convincing in this assay, as would be expected, interpretation of results for compounds with less dramatic activity might be more equivocal if the only results available were from the present assay. One such example is results with vinclozolin, which even at 150mg/kg, only resulted in a significant reduction in epididymal weight with no significant effects on relative seminal vesicle or prostate weight and only a significant elevation in LH levels with no change in testosterone. Nevertheless, within the stated aims of the assay, this compound would still be flagged up for further study. Similar results to vinclozolin were obtained for linuron in the prevalidation studies and inter-laboratory validation exercise and, if changes in thyroid weight and thyroid hormone levels are ignored, then it is only the change in epididymal weight at higher doses of linuron exposure that would flag this compound up as a potential anti-androgen. These particular comparisons also illustrate the limitations of the assay in terms of identifying the MOA, as I am not sure that I would be able to identify an MOA based on the profile obtained for linuron. It may therefore not always be possible to definitively design Tier-2 investigations based on an MOA discerned from the Tier-1 screen using this assay.

3. **Please comment on the biological and toxicological relevance of the assay as related to its stated purpose**

From the pre-validation exercise, a strong foundation has been laid for evaluation and interpretation of results in the assay for compounds for which no information exists about their potential hormone activity. The multiple endpoints of the assay and its relative simplicity mean that its continued application will lead to a progressive ability to categorize chemicals into classes based on their activity profile, even when it is not possible to define a clear MOA. As the profile database expands, so the toxicological utility and predictability of the test is likely to expand also. Because the test uses an intact, adult animal, then compounds may affect target organs or hormone levels via pathways that are unrelated to endocrine disruption *per se*, for example effects on food intake/metabolism that leads secondarily to such changes. This is the 'real world', and it is a strength of the assay that it can integrate such 'biological' effects, though a further reality is that it may be difficult to disentangle such effects from primary endocrine effects in some circumstances (see Q5 below).

4. **Please provide comments on the clarity and conciseness of the protocol in describing the methodology of the assay such that the laboratory can comprehend the objective, conduct the assay, observe and measure prescribed endpoints, compile and prepare data for statistical analyses, and report results**

   Insofar as I feel able to judge (as a scientist running an academic research laboratory), the protocol provided is clearly laid out, understandable and sufficiently detailed to enable an appropriately experienced laboratory to run, complete, evaluate and report results using this assay. There are no major deficits in the protocol that I have spotted, but there are some aspects that could potentially cause confusion. Chief amongst these is the inclusion of hormone assays for FSH, estradiol and to a lesser extent DHT. Although FSH is a key reproductive hormone in the male, its inclusion in the present assay is not especially informative and is not clearly defined. I am uncertain how easy it will prove to interpret treatment-induced changes in FSH levels in the context of the aims of the assay (see comments to Q6), and this may cause confusion to future users of the assay unless its role and significance are better defined (eg its main purpose is to support data for LH to highlight compounds that suppress hypothalamic-pituitary function). The same comments apply to estradiol, as I am unconvinced that our understanding about the role, regulation and significance of changes in estradiol levels in adult male rats is established sufficiently to enable its use in the assay in an informative, as opposed to a confusing/confounding, way. I am not convinced that DHT measurement adds any value to the assay (see comments to Q6).

5. **Please comment on the strengths and/or limitations of the assay**

   The main strength of the assay is that it has a strong foundation based on the pre-validation studies using a variety of compounds with different activities, and the use of multiple endpoints that extend beyond organ weights to include evaluation of hormone levels and how their homeostasis may have altered (at one acute time-point). At the same time, the use of multiple endpoints, and in particular hormone concentrations, raises the possibility of sporadic false (chance) results and identification of false positives. Although these may be weeded out by evaluation of the overall result profile for the compound in question, this may not always be possible, but is probably acceptable as no test will ever be 100% perfect. Of more obvious concern is if there are false negatives. In this regard, it is of interest to consider

the results for linuron from the inter-laboratory validation exercise in some detail, as based on its profile this compound may have come close to being classed as negative, if data for the thyroid axis were excluded from this analysis (and the presumption is that the thyroid changes are secondary to changes in bodyweight and liver weight). My concern is that identification of linuron as a positive compound in Tier-1 depends very much on its effect on epididymal weight and its classification as an anti-androgen is really not possible from the profile obtained, particularly based on results from two of the laboratories involved in the exercise and on comparison of hormone results with those obtained after Phenobarbital treatment (for which no reproductive axis effects would be expected). With linuron, there were no significant changes in testosterone or LH levels in two of the laboratories and no significant changes in relative prostate and accessory sex organ (ASG) weights in two of the laboratories, so it is not obvious that this compound is an anti-androgen. The elevations in FSH levels and in estradiol levels as measured in at least two out of three laboratories suggest that something is going on (though this profile means nothing to me!) but would not class this compound as an anti-androgen. Throughout the report and throughout the inter-laboratory validation exercise, changes in epididymal weight are viewed as evidence of anti-androgenicity, but I am not convinced that this is a logical conclusion. The epididymis is undoubtedly an androgen target organ but this is not nearly so obvious as for prostate and seminal vesicles (weights) and the weight of the adult epididymis is probably determined more by the number of sperm that are present, and stored, in the cauda epididymis (reflecting the completeness of spermatogenesis) than androgen effects *per se*. Though anti-androgens can perturb spermatogenesis, the testis is highly resistant to such effects due to the local, intratesticular production of testosterone. Published studies indicate that linuron may have mixed activity which includes direct effects on the Wolffian duct/epididymis as well as some ability to perturb androgen production, although most such studies have been on the male fetus rather than on the adult as in the present assay. My concern is that when the assay is run in the future in one laboratory for a compound such as linuron, but for which there is no pre-existing data, would it end up as a false negative? I think this is probably unlikely but I use this discussion to illustrate this as a potential limitation of the assay. Overall, I consider that there are sufficient, if inconsistent, changes in the results profile for linuron in the inter-laboratory validation exercise for it to be flagged up for further study for its reproductive effects in Tier-2.

Where the strengths versus weaknesses of this assay are very much in the spotlight is when it comes to analysis of the hormone profile. It is a theoretical strength of this assay that it uses an intact animal ("the real world") in which normal, homeostatic endocrine systems are operating. When any one component of an endocrine loop is disturbed, there should be compensation to bring this axis back to normal levels in terms of biological function. As a consequence, for example, there may be suppression of testosterone production by a compound which then triggers increased LH secretion to act on the Leydig cells to bring testosterone levels back to normal. Measurement of testosterone levels after such an adjustment has occurred may not indicate that anything has happened whereas in fact supranormal LH levels are required to maintain the normal level of testosterone. Such a situation is commonly referred to as "compensated Leydig cell failure". In the pre-validation studies, using compounds with pronounced and established anti-androgenicity, such as flutamide and ketocomazole, such changes in the LH-testosterone axis are highly evident, but this is not the case for much weaker anti-androgenic chemicals such as vinclozolin or, as shown in the inter-laboratory validation exercise, for linuron. If a compensation in LH levels in such situations is relatively minor, this may not be easily discernible against the natural background variation in LH and testosterone levels, which show wide normal fluctuations due to the episodic nature of their secretion. One simple way in which the present assay could be improved to detect such changes would be to determine an LH:testosterone ratio for each individual animal (and then derivation of mean values for the group etc), as this can often indicate that there has been a chronic readjustment of the axis irrespective of what the actual LH and testosterone levels are at any one time in an individual animal; essentially, this ratio is a readout of the current dynamics of the pituitary-Leydig cell axis. This would be a simple refinement to the present study and might enable better identification of such a readjustment in the pituitary-testicular axis, although it is possible that such an adjustment may occur only for a period of time outside of the sampling time used in the assay (see next).

The assay as currently designed involves dosing of animals on day 15 some 2-3 hours prior to euthanisation and sample collection. This is probably a wise choice as it increases the chance of detecting transient effects on hormonal axes that may otherwise not persist to be detectable at a later post-dosing time. The downside of this is that it may detect treatment

effects that are relatively trivial, and are sufficiently transient to have no detectable biological consequence or that it may fail to detect effects that are latent (unless these persist to the following day or 'accumulate'). Such considerations prompt me to conclude that the hormone data should be viewed primarily as playing a supporting role for organ weight/histopathology changes (which provide a summation of effects throughout the treatment period) rather than the other way around. Otherwise, I would be forced to conclude that linuron and phenobarbital have a similar, though not identical, MOA as both mildly suppress testosterone (and possibly LH) levels and elevate estradiol levels, whereas only linuron has any suppressive (anti-androgenic) effects on ASG/epididymis.

It is a fact of life that hormone measurements in the same blood sample in different laboratories can yield dramatically different values for absolute hormone levels, even when all the laboratories are using the same assay kit and procedures. This is well illustrated in the present inter-laboratory validation exercise. Whilst such variation can be taken into account by use of a quality control system, as done presently, this never completely resolves the problem and the bottom line is that it is always difficult to make definitive comparison of absolute hormone levels ***between*** laboratories. There is much more strength when considering changes in absolute hormone levels in different situations ***within*** one laboratory. This is a problem that can be minimized but it will never be completely resolved and the workings of the present in vivo adult rat assay therefore have to take this into account. It is a considerable strength of the adult male assay that it is not reliant on hormone changes *per se*, but on hormone changes in relation to changes in target organ weights.

Some of my reservations about clear identification of linuron as an anti-androgen in the inter-laboratory validation exercise would be removed if absolute ASG weights were used rather than organ weights relative to bodyweight. In this case, linuron would be flagged up as a very clear anti-androgen whereas Phenobarbital would not. However, based on the pre-validation studies by O'Connor using restricted feeding, it was clearly shown that weights of the ASG, other than the epididymis, all declined in parallel with declines in bodyweight. Although these declines were only evident for decreases in bodyweight of 15-25% (again leaving out the thyroid data), this encompasses the magnitude of change in bodyweights for the higher doses of linuron in the inter-laboratory studies. Whilst I understand the basis for using only

relative sex accessory organ weight, I wonder if note will be taken of the absolute organ weights when considering the overall profile and classification of any test compound? Correction of organ weights for changes in bodyweight will help minimize identification of false positives, but my concern would be that it may also result in false negatives as might nearly have happened for linuron in at least one of the laboratory studies. I am not certain that it can be concluded definitively that a decrease in bodyweight will always lead to a secondary reduction in ASG weight, irrespective of the mechanism of action that initially precipitated the reduction in bodyweight; it also needs to be remembered that reductions in testosterone levels may itself result in loss of bodyweight/altered body composition which is a potentially confounding effect (though may not be too important in this relatively short assay).

An undoubted strength of the assay in terms of its multiple endpoints is that it also has the potential to identify compounds which have reproductive or thyroid target organ effects but which do not operate primarily through endocrine effects. It is possible that linuron may be a somewhat unclean example of this as it may have direct effects on the epididymis as well as endocrine effects directly on the testis. However, in the context of development of the present Tier-1 assay, it is unclear to me if such compounds would be flagged up for further study if there is no evidence for any endocrine activity.

6. Please provide comments on the impacts of the choice of:

a. test substances

The pre-validation studies have used compounds with a wide range of hormonal or other activities and these have provided a robust evaluation of the effectiveness of the assay and of its sensitivity and discriminatory powers. Much of this data has been obtained only in single laboratories and with considerable experience of running this assay. This may not provide an accurate guide as to how usable the assay will be when let loose in the "real world". However, these studies have served their undoubted purpose. In terms of the inter-laboratory validation exercise, I endorse the selection of linuron and phenobarbital as test compounds, as neither has profound endocrine disruptor activity comparable to a chemical such as flutamide, for example. Phenobarbital has quite major biological effects and has effects on bodyweight and the thyroid axis but was not expected to impact on the reproductive axis. It

therefore provided a good choice via which to see how discriminating the assay could be in picking up thyroid changes whilst showing no effect on the reproductive axis. This was largely achieved. Similarly, selection of linuron was a good choice because its use in the assay in at least three different laboratories in the pre-validation exercise had shown reproducible effects on androgen target organs and on hormone levels but only at high doses and only in a rather selective way; again, it did not have profound activity such as a compound like flutamide. Its inclusion in the inter-laboratory validation exercise was therefore a good choice as it has the sort of activity that would make it a good candidate for becoming a "false negative". The fact that all three laboratories provided statistically significant evidence for its 'anti-androgenicity' (based on effects on epididymal weight at one or more doses) is therefore reassuring, though I mention elsewhere my concerns about the use of epididymal weight as a definitive measure of anti-androgenicity.

b. analytical methods

Essentially two analytical methods are used as part of the test, hormone assays and selective evaluation of organ histopathology (testes, epididymides, thyroid). If the test is to be practicable and applicable in laboratories around the world, it demands consistency in terms of assay kits used as there is enough variation anyway in hormone measurements between laboratories when using the same assay kit. Standardization of the kits used and consequently of the method used is therefore an important step towards uniformity as well as minimizing inter-laboratory variation. However, such variation is commonplace and likely to be considerable when, and if, the assay is put into widespread use by laboratories that have little experience with running hormone assays. For this reason, organ histopathology will continue to be an important component of the test as it may provide confirmation of a target organ effect for a compound with relatively limited activity. It is perhaps not reassuring that no histopathology was picked up for linuron in any of the laboratories except for one laboratory reporting very minor testicular changes. As all of the laboratories involved are experienced in organ histopathology, it suggests that this assay will only be useable in laboratories with resident histopathology expertise and is likely to be insensitive on its own.

I am not certain of the relevance and importance of FSH measurements in the current assay in relation to its overall purpose. FSH levels will only normally increase significantly when

there is quite severe impairment of spermatogenesis such that testicular weight is decreased and secretion of inhibin-B is reduced. Therefore there is no obvious benefit of measuring FSH in this situation when the information will already be provided by a more easily measurable endpoint ie. testis weight. In the inter-laboratory validation exercise, linuron exposure mildly elevated FSH levels whereas phenobarbital mildly decreased FSH levels and in neither case was it obvious why this should have occurred due to an anti-androgenic mechanism or due to any effects on the testis itself, which either did not occur (phenobarbital) or were trivial in nature (linuron). Estradiol is a negative regulator of FSH, but levels were significantly elevated in both linuron and phenobarbital-exposed animals, which thus provides no consistent explanation for the altered FSH levels. This also draws into focus why estradiol levels should be increased (there is no obvious explanation for either treatment) and what is the precise importance of estradiol measurements in the current assay? I am not sure that we yet fully understand the roles of estradiol in the male and this may make it difficult to interpret changes in estradiol levels, as for example in the inter-laboratory comparison, and what this may mean in terms of the endocrine disrupting potential of a test compound. Additionally, blood levels of estradiol are very low in an adult male rat, are not easy to measure by assay (the assay used presently has a high coefficient of variation) and is likely to prove one of the more problematical measurements once the assay is adopted by a wider number of laboratories, especially those with little experience in running hormone assays.

I am unconvinced of the need for measurement of DHT in the present assay. The only obvious benefit of its inclusion is that it may help to identify the MOA for compounds that act as 5α reductase inhibitors. However, as such compounds should also be picked up by their effects on weight of androgen target organs, inclusion of this particular hormone assay in this Tier-1 screen is probably an unnecessary complication unless experience subsequently proves that 5α reductase inhibition is a common effect of compounds under investigation.

c. statistical methods in terms of demonstrating the performance of the assay

As far as I am able to judge, the statistical methods used for analysis of the significance of effects, for analysis of trends and for comparison of variability in methodology between laboratories is appropriate. However, I am not sure that any statistical package can truly

evaluate the performance of the assay as this has to integrate all of the organ and hormonal data in a way that allows objective decision making and classification and I am not certain that this is possible. Instead, I feel that such decision making will be based not on appropriate individual statistical tests but analysis of the data by experts who have experience of the test and with results and variability in responses that it shows for different chemicals.

7. **Please provide comments on repeatability and reproducibility of the results obtained with the assay, considering the variability inherent in the biological and chemical test methods**

From the published studies and reports of the pre-validation investigations, the reproducibility of the assay appears to be very good. Nevertheless this has only been investigated for one or two compounds and it would be a major surprise if the effects of a potent anti-androgen such as flutamide, for example, were not picked up robustly in the assay when used by different laboratories. It is beyond dispute that the assay works effectively in the various test situations in which it has been applied, but the challenge remaining is to decide how it will be operated in practice by different laboratories in the future and what will be the false negative and false positive rates. Will "linuron-like" compounds always be positively identified? In addressing this question for linuron itself, it is reassuring that its effects have been picked up in the assay by different laboratories both in the pre-validation studies and in the inter-laboratory validation exercise. This imparts a degree of confidence that similar compounds, for which the activity is completely unknown, will be picked up during future use of this assay by different laboratories. Furthermore, the studies with phenobarbital reinforce the view that the assay is robust, as it was negative in all of the laboratories for reproductive organ endpoints whilst showing positive activity for thyroid gland weight and thyroid hormone changes. Nevertheless, it is somewhat disconcerting that all three laboratories identified significant effects on blood levels of FSH and estradiol after phenobarbital exposure, the changes with FSH even being observed at the lowest treatment dose. The most rational explanation for this observation is that it is the elevation in estradiol levels that is responsible for the suppression of FSH, but the question then is whether phenobarbital would be classified as an endocrine disruptor of the reproductive system based on these observations? I am not entirely sure what the purpose of inclusion of FSH and estradiol measurements is in the current assay, and this is a point that is discussed elsewhere.

The changes in thyroid gland weight and thyroid hormone levels appear particularly robust in the pre-validation studies and especially in the inter-laboratory validation exercise. I do not feel sufficiently expert to comment in any informed way on the value of the assay as applied to the thyroid axis, but I am disappointed to note that only one of the laboratories was able to confirm the thyroid hormone changes with appropriate histopathological changes; one of the other laboratories also noted minor changes but only at the very highest dose. This questions how useful target organ histopathology is going to prove in this assay, when it cannot match up with the hormone data which is changing in such a reproducible fashion.

3.5 Thomas Zoeller Review Comments

REVIEW:

VALIDATION OF 15-DAY INTACT ADULT MALE RAT ASSAY AS A POTENTIAL SCREEN IN THE ENDOCRINE DISRUPTOR SCREENING PROGRAM TIER-1 BATTERY

Introduction

Section 408(p) of the Federal Food Drug and Cosmetic Act (FFDCA) requires the U.S. Environmental Protection Agency (EPA) to: *develop a screening program, using appropriate validated test systems and other scientifically relevant information, to determine whether certain substances may have an effect in humans that is similar to an effect produced by a naturally occurring estrogen, or other such endocrine effect as the Administrator may designate [U.S.C. a(p)].* The 15-day intact adult male rat assay as an alternate component of the Tier-1 screening battery was recommended by the EDSTAC committee and has been developed by industry in the intervening years. The current document represents a considerable amount of effort focused on evaluating the ability of this assay to identify chemicals that interfere with the androgen and thyroid systems. In general, an environmental endocrine disruptor is defined as *an exogenous agent that interferes with the synthesis, secretion, transport, binding, action or elimination of natural hormones in the body that are responsible for the maintenance of homeostasis, reproduction, development, and/or behavior*.

In general, this is an ambitious project that was not managed by EPA in a manner required to achieve the stated goals. This is unfortunate. There are four categories of weaknesses, each of which was preventable. These include a) lack of performance standards and criteria for RIAs, b) failure to develop a logical framework in which to interpret the results *a priori*, c) failure to carefully control contents of the feed and determine the degree to which this affects the performance of the assay, d) failure to carefully inspect the data generated. Each of these categories is discussed in greater detail below. However, not all of these categories fit neatly into the charge questions; therefore, I will discuss these in greater detail here.

Performance standards and criteria of the RIAs. The RIA data provided in this document show a great deal of variability in hormone levels of the control animals across laboratories. However, it is not possible to identify the source of this variation as being technical or biological because the types of studies required to separate these two sources of variation were not performed. Specifically, the EPA should develop and distribute, or should contract to develop and distribute, the quality control standards to all laboratories performing RIAs in the commission of the EDSP. These centralized standards would greatly decrease the variance across laboratories and would enhance the reliability of the assays. In addition, the three laboratories used different commercial kits for the various RIAs and EPA did not require that the RIAs were validated (in the case of heterologous assays) or that the QC was performed as described by the kit manufacturer or that the performance fell within the range defined by the manufacturer. There is no question that these problems can account for a great deal of variability in the RIA results, and that a minimal amount of thought and effort by the EPA at the beginning of this project could have prevented it. It must be remembered that RIAs have been in use for nearly 50 years, and methods for validating assays and standardizing them across laboratories have been very well developed.

Because of these technical problems, the degree of biological variability in hormone levels and effects of treatments on hormone levels, cannot be ascertained. Certainly, some of the variability observed in this exercise is related to biological variability. One can imagine a number of differences among housing conditions that could account for this. For example, the feed and animal housing in use in the EDSP was not well controlled. We know that there is much greater variability in the contents of the feed than appears on certificates from the suppliers (1, 2). Differences in the amount of isoflavones in our experiments can make at least a

50% differences in the concentration of total T_4 in serum. Important constituents include not only isoflavones that can act as estrogens and thyroid peroxidase inhibitors, but also iodine, which can greatly influence thyroid function. The EPA made two logical mistakes in the way they present the criteria for the feed. The first paradox is that they argue that 15 days of a specific feed is not long enough to have significant impact on hormone levels or on the response to treatments (without supporting evidence). However, if this is true, then the feed the animals were provided prior to the beginning of the experiment is more likely to have an impact on the experiment, but this is not specified. Controlling the components of the feed will doubtlessly be difficult. However, for EPA to state that, "Certified animal feed will be used, guaranteed by the manufacturer to meet specified nutritional requirements. Analysis will include ensuring that heavy metals, pesticides, and phytoestrogens (e.g., genistein, daidzein, and glycitein) are not present at concentrations that would be expected to affect the outcome of the study", (Appendix C, page 6 of 21) provides no guidance to a laboratory trying to perform this assay to the best of their ability. EPA has not cited information about the effects of phytoestrogens in the feed and the consequences on "the outcome of the study".

Failure to develop a logical framework in which to interpret the results a priori. The EPA document describes in the introductory material (page 4, *Test Development*) that detailed review papers are used as the basis of the test. However, this does not appear to be the case. The review material used do not provide the EPA with a specific framework in which to predict the kinds of effects that would be observed in the 15-day adult male assay. A case in point is the affect of Linuron on the HPT axis. The data presented in this document show that Linuron can produce a significant (and robust) decrease in serum total T_4, but that the thyroid gland and serum TSH is only slightly – or not – affected. Therefore, the serum T_4 levels are considered to be uninformative. This interpretation is supported by the observation that many (27) of the 29 chemicals evaluated in this assay can also cause a decrease in serum total T_4.

This is a highly uninsightful interpretation and reflects that lack of forethought put into the interpretation of possible results. First, it is illogical to base an interpretation on the proportion of chemicals that reduce serum total T_4 in a series of "prevalidation" studies. These chemicals were selected because of preliminary evidence that they are endocrine disrupting compounds. Might they considered a non random sample of chemicals? Second, because the EPA failed to develop endpoints of thyroid hormone action in the 15-day intact adult male assay,

the assay itself is asymmetric; that is, there are endpoints of androgen action (organ weight and histopathology), but not of thyroid hormone action. Thus, the assay itself is capable of identifying an antiandrogen that causes a reduction in serum testosterone but does not increase LH, but is not capable of identifying an anti-thyroid agent similarly. The EPA's current interpretation would likely eliminate PCBs as anti-thyroid agents. Although some studies have shown that PCBs can cause an increase in serum TSH, many show that PCBs do not increase TSH levels. Thus, this profile would look like the effects of Linuron and would be ignored. The EPA authors do not explain why two chemicals (Linuron and Phenobarbital) that act by the same mechanism (increase liver clearance of T_4) can have two different effects on serum TSH. To what extent must TSH levels be increased before there are measurable changes in thyroid weight and histopathology? These issues should have been discussed prior to the commission of this assay for inter-laboratory validation and potential solutions identified.

Failure to carefully control contents of the feed and determine the degree to which this affects the performance of the assay. To be sure, this is a difficult task. NIEHS recently sponsored a workshop on animal feed in EDC research and included manufacturers of animal feed. This EPA document ignores the importance of this issue except to state that the level of phytoestrogens should be below that "expected" to interfere with the performance of the assay. In addition, many of these isoflavones inhibit thyroperoxidase and, in our lab, the presence/absence of soy protein in the feed can alter thyroid hormone levels very significantly. Thus, different diets will interact in this assay in a way that increases the biological variability.

Failure to carefully inspect the data generated. Table one is a compilation of mean±SEM for testosterone, LH, T_4, T_3, and TSH. These data were recruited from the individual reports of the 3 laboratories. Highlighted are data cells that contain exactly the same SEM value (to 3 decimal places). For example, T_3 levels in the Linuron-treated groups (0, 50, 100, mg/kg) reported in the WIL report have an SEM of 2.859. Moreover, this value in the Phenobarbital treatment groups is exactly the same. It would appear to be highly unlikely that the standard error of the mean, with 15 animals/treatment group, is exactly the same in all of these groups.

The problems outlined above and described below render this inter-laboratory exercise incapable of being interpreted. It is difficult not to conclude that EPA has not lived up to their charge to validate this assay and the produce a credible document.

RESPONSES TO CHARGE QUESTIONS

1. Please comment on the clarity of the stated purpose of the assay.

The stated purpose of the assay is perfectly clear. A point of confusion though is the relationship between validation of individual assays and validation of the battery. Tier-1 and Tier-2 batteries are complex, and for them to be informative as envisioned, each of the component assays must be reliable and their interpretation must be guided within the context of the tier itself. However, the discussion in the document does not clarify the relationship between validation of the 15-day adult male assay and the Tier-1 battery itself.

2. Please comment on the clarity, comprehensiveness and consistency of the data interpretation with the stated purpose of the assay.

The manuscript clearly describes the logic used to interpret the data provided by the 3 laboratories. The methods employed and the endpoints collected are clear. The EPA document, and the individual reports from RTI, WIL and Charles River, indicates that because the RIAs are so variable both within and between laboratories, the hormone levels are to be used for supportive evidence for a role of a chemical as an endocrine disruptor (androgen or thyroid), but that body and organ weight and histopathology should represent primary data. Thus, the endpoints captured including body and organ weight and histopathology (thyroid, testes, epididymides), provide primary information about the toxicity of a chemical and the MOA as an endocrine disruptor. There are two problems with this logic. First, the tissues employed as endpoints of androgen and thyroid disruption represent endpoints of androgen action (e.g., epididymus, seminal vesicles), but there are no endpoints of thyroid hormone action that would be equivalent to epididymus or seminal vesicles. Thus, chemicals like linuron that can reduce circulating levels of thyroid hormone without affecting (or perhaps even lowering) serum TSH may not produce an effect on the thyroid gland itself (through elevated TSH) and will therefore be ignored. Thus, there is a fundamental flaw in the endpoints designed for capture in this assay. Second, although a considerable problem is that of the high variability in hormone levels both

within and across laboratories, there may be a solution to this problem (see below). In the absence of providing reliable data for hormone levels, this and the other in vivo assays will be severely compromised.

The relationship between body weight reductions produced by toxicity or by caloric restriction is a complex one and the background information provided is interesting and important. Briefly, these data show to what extent total body weight must be reduced (caused by caloric restriction) before impacting the weight of the various organs or hormone levels. This information is used in the interpretation of the data arising from toxicant treatments by assuming that the relationship between total body weight and organ weight will hold for all toxicants. Caloric restriction is known to produce a significant and potent reduction in serum thyroid hormone levels, which can be blocked by placing lesions in the hippocampus (3, 4). Thus, the fasting-induced reduction in thyroid function is mediated by the central nervous system. In addition, this effect also involves the type 2 deiodinase (5, 6). Therefore, the effect of caloric restriction on the HPT axis is centrally mediated and may respond to toxicants in ways that do not simply duplicate caloric restriction. Perhaps changes in the use of specific metabolic fuels (fat, protein, carbohydrate) can elicit this response in the absence of large changes in body weight. In contrast, perhaps some chemicals can block this effect regardless of body weight changes? Although somewhat speculative, this hypothesis is clearly plausible and the simple assumption that body weight will always be related to organ weight in a particular way seems both unnecessary and dangerous.

3. Please comment on the biological and toxicological relevance of the assay as related to its stated purpose.

Section 3.1 discusses the relevance of the bioassay. This section begins with statements about how "Numerous EACs (one negative and 28 positive test chemicals)…" have been tested in the 15-day intact adult male assay, but the data are not presented nor are they fully referenced. In addition, an examination of Table 4 lists these chemicals with a very cursory description of their MOA. For example, the document states that, "Thus, throughout prevalidation, the intact adult male assay has been run with 29 different test chemicals at various times in six different laboratories (four chemical industry laboratories and two different contract research organizations, or CRO laboratories). In some instances the same chemicals were tested

in more than one laboratory at different times as shown in Table 4." This is a very misleading statement that is not supported by the information presented in Table 4 nor is it supported by the discussion in Section 3. Therefore, it undermines the credibility of the current document more than it supports the strength and validity of the 15-day adult male assay.

In addition, it is not clear from the remainder of this section why this information is being presented. This section could have provided a logical basis for the design of the assay. To do so, it would have to review the basic endocrinology of the androgen and thyroid systems to the extent that the biological and toxicological relevance of a 15-day assay performed in the adult male would be supported. However, no such discussion is presented, and the choice of endpoints identified in the 15-day adult male assay lack clear support, which becomes apparent when the data are interpreted. Section 3.1 presents a very cursory review of a variety of chemicals evaluated in a variety of experimental designs. This is a highly confusing section that does not advance arguments in support of the assay. Worse, this section indicates that a variety of chemicals with known endocrine activities were evaluated in the 15-day adult male assay. Thus, it gives the impression that chemicals were defined as having endocrine activity in other kinds of experimental designs and then evaluated in this 15 day assay. If this assay has provided fundamental new information about the endocrine activity of various chemicals, this section does not provide a credible review of this. Overall, this section is a pivotal section that fails to provide a careful review of the literature that is relevant to the 15-day adult male assay, nor does it provide a convincing argument that supports the expectation that this assay could provide information about the biology of the androgen or thyroid system. Indeed, it is difficult to imagine that new insight into these endocrine systems will be generated from such an experimental paradigm.

Finally, the toxicological relevance of this assay also is not made clear in this section. This is not to say that the assay does not have toxicological relevance. Rather, the information required to conclude that this assay has toxicological relevance is simply not presented. Specifically, the document begins with a definition of an "endocrine disruptor". This definition, copied verbatim above, should provide the foundation of section 3.1. That is, if a chemical acts as an EDC (by definition), then one can make predictions about the endpoints that should reveal this and can be used to identify EDCs. The form of this section is not such that this logic is presented.

Section 3.2 also is an important part of the document that could provide a logical framework for the design and interpretation of the 15-day adult male assay. However, this section is written in such a way that it fails to provide credibility to the overall document. For example, this section introduces the concept of chemically-responsive "fingerprints" that may provide information about the mode of action of a specific chemical. This concept of "fingerprints" turns out to simply mean that if endpoints at different levels of the HPG and HPT axis are evaluated, one may obtain basic information about the site at which a chemical interferes with endocrine activity. As an example, work attributed to O'Connor et al. (1998a and 2000c) is provided. This work apparently showed that, "Correspondingly, the ability of flutamide to block the negative feedback effect of testosterone and DHT at the hypothalamic and pituitary levels resulted in the secretion of gonadotropin-releasing hormone (GnRH) and LH, respectively, and the subsequent production of testosterone by the Leydig cells of the testes. Thus, the chemical-responsive "fingerprint" of an AR antagonist such as flutamide is a decrease in ASG weight and increased serum concentrations of testosterone and LH." This is exactly the kind of logic that should have been presented in section 3.1 and the basis for that logic is basic endocrinology. However, the references that appear to be used to support this statement do not, in fact, provide support. Neither of these citations measured – or even mentioned – GnRH, and one (O'Connor et al., 2000c) does not report studies of flutamide. Moreover, even a cursory knowledge of research on GnRH would have informed the writers of this section that GnRH secretion is a very technically demanding endpoint to capture, and there are not many laboratories with the skill or equipment to perform such studies. Thus, to make a statement such as that cited above without the proper support undermines the credibility of the document in providing a logical framework upon which this 15-day adult male assay is developed.

There are two important points that this section (3.2) illustrates. First, the concept that identifying "fingerprints" of endocrine activity as a novel approach requires that one suspend decades of basic research in endocrinology that informs such an approach. Essentially, this "fingerprint" simply means that one may infer the site within an endocrine axis that a chemical acts to interfere with the system by simultaneously capturing endpoints at different levels within the axis (e.g., gonadal and pituitary hormones). It is reasonable that the 15-day adult male assay for EDCs be placed within an endocrinological context to have credibility both as an individual assay and as a component of tier-1 screens. However, the endocrinological

context is not provided in this document and the writers appear to be unaware of this context. This undermines the presentation of the data and its interpretation later in the document.

4. Please provide comments on the clarity and conciseness of the protocol in describing the methodology of the assay such that the laboratory can:

This question appears to refer to Appendix C. Thus, the answers below are focused on this section.

a. comprehend the objective. The objective of the 15-day intact adult male assay is to contribute to the first tier of screens for EDCs. Thus, it is intended to identify *new* chemicals (i.e., chemicals for which little information is available) that interfere with estrogen, androgen or thyroid activity. This much is clear.

b. conduct the assay. The methods described appear to be sufficient to guide an independent laboratory to conduct the assay.

c. observe and measure prescribed endpoints. The information provided is sufficient in most cases. However, as described more completely below, the EPA should provide additional guidance and criteria for helping independent laboratories perform hormone analysis.

d. compile and prepare data for statistical analyses. The information provided is sufficient.

e. report results. The information is sufficient.

5. Please comment on the strengths and/or limitations of the assay.

This question can be addressed within the context of section 2.3.1 of the document. These identified strengths are as follows, and

• ***Allows for a high-order neuroendocrine assessment of male reproductive and thyroid function due to the use of an intact endocrine system (i.e., HPG and HPT axes).*** The *in vivo* nature of this assay means that the interactions of hormone signaling within the HPG and HPT axes can be captured, and this is a genuine strength of the assay. If this is what is meant by "high-order neuroendocrine", then I agree. However, the role of the hypothalamus in mediating

effects of chemical treatment on the endpoints captured in this assay cannot be ascertained. Thus, the term "neuroendocrine" is overstated at best.

• ***Advances scientific understanding through its MOA and, perhaps, mechanistic approach (i.e., measurement of serum concentrations of reproductive steroids, gonadotropins and thyroid hormones).*** This is an overstatement at best. It should be clear that the 15 day adult male assay cannot be considered a "mechanistic" approach to understanding new information about basic endocrinology. At best, this assay can identify a broad range of chemicals that interfere with androgen and thyroid endocrine system.

• ***Provides MOA data (e.g., differentiates between receptor and nonreceptor-mediated effects) that can be used to tailor the design of more definitive Tier-2 tests to focus on selective endpoints to accurately identify potential hazards, define dose responses, and determine the level of risk of potential endocrine disruptors.*** The 15-day intact adult male assay cannot differentiate between "receptor and nonreceptor-mediated effects" and it doesn't need to. In fact, which receptor is being addressed in this statement" the androgen, estrogen or thyroid hormone receptor? The FSH, TSH or LH receptor? Unfortunately, this statement is so naïve that it undermines the credibility of this section. The goal of the 15-day intact adult male assay is not to determine the mechanism of action, but to recruit information about the ability of a chemical to act as an EDC on the androgen or thyroid system. The best that it can hope to do is to help to reduce the number of false-negatives in the Tier-1 screen.

• ***Allows for the maximum tolerated dose (MTD) to be readily defined since mature animals are less susceptible to marked changes in growth and less susceptible to nonspecific alterations in endpoints secondary to bodyweight changes.*** This statement, taken literally, states that a strength of this assay is that adult animals are less sensitive to the toxic effects of chemicals than less mature animals. Is this a true strength? The document does not defend such a statement in any manner, which would seem to be required for such a statement.

• ***Flexible for modifying or adding apical, histological and hormonal endpoints in the context of a single assay to detect other potential endocrine-related effects as future application may dictate.*** This is a potential strength, but the methods by which an additional endpoint would be validated is not clear. Moreover, the methods by which the endpoints described in the current version of the assay are not well supported in this document or in the supporting publications.

From an endocrinological perspective, important questions about the sensitivity of these endpoints to specific disruptions in endocrine action are unanswered.

• ***Complies with the basic principles of good laboratory animal practice (i.e., three R's - Reduce, Refine, and Replace), specifically through the effective use of a minimal number of animals.*** This is a clear strength of the assay. It is clear that a single in vivo assay can test the ability of unknown chemicals to interfere with endocrine action at a number of levels at once. This could not be accomplished by a single – or even multiple – in vitro assays.

• ***Complies with the expected simplicity and rapidity of a screen prescribed by the EDSTAC since the in-life portion of the assay is readily applied and minimal in duration.*** In principle, this assay should provide rapid information about the ability of a chemical to interfere with endocrine action. The degree to which this is a strength is related to the ability of the EPA to sharpen this assay, its justification, commission and interpretation.

Assay Limitations. The limitation of the assay is that the background information does not provide essential information required to interpret the results. For example, to what extent must TSH be elevated to produce histological changes in the thyroid gland? To what extent must TSH be elevated to produce thyroid tumors in the SD rat? How can total T_4 be reduced without changing serum TSH? Why do some chemicals cause a reduction in total serum T_4 without a concomitant change in serum TSH, yet other chemical produce a change in serum TSH?

6. Please provide comments on the impacts of the choice of:

a. test substances. There were a large number of chemicals that could have been used, but this does not seem to be an essential issue. It is not clear why Phenobarbital and Linuron were used because both activate liver enzymes that likely cause a reduction in serum thyroid hormone, but these chemicals are as suitable as other chemicals that interfere with androgen or thyroid action. Interestingly, these chemicals illustrate a serious weakness both in the commission of the studies and in the interpretation of results. Specifically, the conclusions are based on the expected findings and not on a logical framework that was established *a priori*. This issue will be addressed elsewhere.

b. analytical methods. EPA made a serious mistake in not setting very precise performance standards for specific methods, especially the RIA. This is the most important reason that the data appear to indicate that hormone levels are variable from one lab to another. Considering that RIAs have been in existence for nearly 50 years, it is remarkable that EPA set up such an inter-laboratory test with no apparent thought with the ways other entities (e.g., CDC) have design to allow quantitative comparisons of results between various independent laboratories. These issues are described below.

1. Radioimmunoassay (RIA). Minimally, EPA must establish performance standards for the RIA that a laboratory uses in the commission of this (and all other in vivo) assay. These performance standards should include:

i. Intra-assay variation. Each of the commercial kits reported by the laboratories in this study reported an intra-assay variation far below that reported by the laboratory using the kit. What reason could there be for EPA to accept data that do not meet the performance standard of the kit? Thus, EPA should specify that the end user of the kit is at least using the kit properly as defined in part by their reported intra-assay variation. In fact, the reported difference in performance of the assay may be related to the difference in the way in which the intra-assay variation was defined. Companies using these kits should use the same method to define intra-assay variation as that described in the kit. If they do not, they are not actually measuring the intra-assay variation. Related to this, in the individual description of the work, one company specified that all samples from a single experiment were run within a single assay. Another company did not specify and the third reported that when all samples could not be run in a single assay, the samples were distributed across groups in different assays. This is simply unacceptable and EPA should require that all samples be run in a single assay.

ii. Inter-assay variation. EPA should require that companies measure inter-assay variation as it is described in the kit they are using, and that this measure falls within the limits reported by the kit manufacturer.

iii. Validation. EPA should either require that companies use homologous assay kits (i.e., designed and validated for rat), or require that companies validate the use of heterologous assays in their own lab. For example, we have demonstrated in my lab that the kit designed to measure total T_4 in humans is not valid for rat serum. Thus, these kits do not always perform

well for other species and this is a minimal requirement that EPA should have anticipated. In addition, EPA should specify the validation method. This would likely include a linearity check using dilutions of rat serum in addition to "spiking" rat serum with the hormone of interest and evaluating "recovery".

iv. *Inter-laboratory consistency.* Consistency across laboratories could be better standardized if EPA developed and distributed the QC standards to which all commercial kits would have to be cross-calibrated. This is an issue that should easily have been anticipated. The Centers for Disease Control and Prevention will have a system that that the EPA could model on a smaller scale. They likely have a set of performance standards for contract (clinical) laboratories that will be very helpful.

c. ***statistical methods in terms of demonstrating the performance of the assay.*** I was honestly surprised to see that the statistical methods used to characterize the performance of the assay across laboratories were almost completely designed post-hoc. Perhaps the degree of variation was higher than anyone expected. Certainly, variation in hormone levels among controls could be reduced considerably by the methods described above. However, variation in treatment effects could also be from differences in feed. While the same feed was used for this 15-day study, is it possible that the animals were not on the same feed prior to this? Thus, if, as the EPA speculate, phytoestrogens in the feed would not impact adult males in 15 days, it may also follow that the feed the animals were on would affect the results when the animals were only taken off that feed for 15 days. The point here is that so many mistakes were made in standardizing the assay across laboratories that it is doubtful that finding just the right statistical analysis will help in this particular case.

Please provide comments on repeatability and reproducibility of the results obtained with the assay, considering the variability inherent in the biological and chemical test methods. I do not believe that the variability observed in the assay is inherent in the biological and chemical test methods. Unfortunately, this exercise has failed to answer this question.

REFERENCES

1. **vom Saal FS, Richter CA, Mao J, Welshons WV 2**005 Commercial animal feed: variability in estrogenic activity and effects on body weight in mice. Birth Defects Res A Clin Mol Teratol 73:474-475

2. **Vom Saal FS, Richter CA, Ruhlen RR, Nagel SC, Timms BG, Welshons WV** 2005 The importance of appropriate controls, animal feed, and animal models in interpreting results from low-dose studies of bisphenol A. Birth Defects Res A Clin Mol Teratol 73:140-145

3. **Blake NG, Johnson MR, Eckland DJ, Foster OJ, Lightman SL** 1992 Effect of food deprivation and altered thyroid status on the hypothalamic-pituitary-thyroid axis in the rat. J Endocrinol 133:183-188

4. **Blake NG, Eckland DJ, Foster OJ, Lightman SL** 1991 Inhibition of hypothalamic thyrotropin-releasing hormone messenger ribonucleic acid during food deprivation. Endocrinology 129:2714-2718

5. **Coppola A, Hughes J, Esposito E, Schiavo L, Meli R, Diano S** 2005 Suppression of hypothalamic deiodinase type II activity blunts TRH mRNA decline during fasting. FEBS Lett 579:4654-4658

6. **Diano S, Naftolin F, Goglia F, Horvath TL** 1998 Fasting-induced increase in type II iodothyronine deiodinase activity and messenger ribonucleic acid levels is not reversed by thyroxine in the rat hypothalamus. Endocrinology 139:2879-2884

Table 3-1. Tom Zoeller: Effect of Linuron on Selected Hormones in the three studies.					
	Site	**Control**	**50mg/kg**	**100 mg/kg**	**150 mg/kg**
Testosterone (ng/mL)	WIL	6.137 ±0.95	6.317 ± 1.267	6.225 ± 1.267	4.102 ± 1.312
	RTI	3.37 ± 0.75	4.07 ± 0.42	2.56 ± 0.42	2.49 ± 0.43
	Charles-River	9.93 ± 1.873	4.83 ± 0.960	3.98 ± 1.065	3.28 ± 0.641
LH (ng/mL)	WIL	0.693 ± 0.098	0.880 ± 0.098	0.667 ± 0.098	0.629 ± 0.102
	RTI	1.28 ± 0.08	1.33 ± 0.07	1.31 ± 0.08	1.41 ± 0.08
	Charles-River	2.18 ± 0.127	1.75 ± 0.095	1.78 ± 0.137	1.86 ± 0.179
T_4 (µg/dL)	WIL	4.973 ± 0.217	2.687 ± 0.229	1.287 ± 0.205	0.521 ± 0.064
	RTI	5.55 ± 0.16	3.87 ± 0.16	2.61 ± 0.16	1.69 ± 0.17
	Charles-River	4.73 ± 0.151	3.10 ± 0.151	1.82 ± 0.151	1.54 ± 0.151
T_3 (ng/dL)	WIL	79.993 ± 2.859	78.627 ± 2.859	69.873 ± 2.859	66.943 ± 2.96
	RTI	87.53 ± 2.92	85.66 ± 2.92	76.90 ± 2.92	77.98 ± 3.02
	Charles-River	81.65 ± 2.593	64.85 ± 2.593	65.41 ± 2.593	56.15 ± 2.684
TSH (ng/mL)	WIL	15.393 ± 1.697	13.947 ± 1.357	12.247 ± 1.357	11.386 ± 1.41
	RTI	18.53 ± 3.18	18.51 ± 1.80	11.90 ± 0.88	13.85 ± 1.53
	Charles-River	13.10 ± 1.681	23.35 ± 2.877	25.74 ± 2.024	29.73 ± 2.269

Table 3-2. Tom Zoeller: Effect of Phenobarbital on Selected Hormones in the three studies.					
	Site	**Control**	**25 mg/kg**	**50 mg/kg**	**100 mg/kg**
Testosterone (ng/mL)	WIL	6.137 ±0.95	4.475 ± 0.674	3.501 ± 0.674	2.870 ± 0.698
	RTI	3.37 ± 0.75	2.62 ± 0.52	2.64 ± 0.52	2.66 ± 0.56
	Charles-River	9.93 ± 1.873	4.83 ± 0.960	3.98 ± 1.065	3.28 ± 0.641
LH (ng/mL)	WIL	0.693 ± 0.098	0.647 ± 0.098	0.560 ± 0.098	0.779 ± 0.102
	RTI	1.28 ± 0.08	1.28 ± 0.07	1.24 ± 0.08	1.18 ± 0.08
	Charles-River	2.18 ± 0.127	1.81 ± 0.098	1.44 ± 0.054	1.56 ± 0.068
T_4 (µg/dL)	WIL	4.973 ± 0.217	4.133 ± 0.293	2.893 ± 0.148	1.707 ± 0.204
	RTI	5.55 ± 0.16	4.66 ± 0.16	4.32 ± 0.16	2.97 ± 0.17
	Charles-River	4.73 ± 0.151	3.75 ± 0.151	3.64 ± 0.151	2.62 ± 0.156
T_3 (ng/dL)	WIL	79.993 ± 2.859	72.040 ± 2.859	62.773 ± 2.859	57.086 ± 2.96
	RTI	87.53 ± 2.92	78.24 ± 2.92	73.28 ± 2.92	58.14 ± 3.14
	Charles-River	81.65 ± 2.593	64.85 ± 2.593	65.41 ± 2.593	56.15 ± 2.684
TSH (ng/mL)	WIL	15.393 ± 1.697	24.013 ± 2.816	25.833 ± 2.816	31.857 ± 2.92
	RTI	18.53 ± 3.18	21.69 ± 1.79	26.6 ± 2.42	26.38 ± 3.50
	Charles-River	13.10 ± 1.681	23.35 ± 2.877	25.74 ± 2.024	29.73 ± 2.269

Appendix A

CHARGE TO PEER REVIEWERS

CHARGE TO PEER REVIEWERS

for

INDEPENDENT PEER REVIEW OF THE 15-DAY INTACT ADULT MALE RAT ASSAY AS A POTENTIAL SCREEN IN THE ENDOCRINE DISRUPTOR SCREENING PROGRAM (EDSP) TIER-1 BATTERY

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 15-day intact adult male rat assay. The intact adult male assay consists of multiple endpoints; principally, terminal weights of primary and secondary sex organs and thyroid gland, histology of the testes, epididymides and thyroid, and serum concentrations of reproductive steroids, gonadotropins and thyroid hormones.

According to numerous reports published in peer-reviewed scientific journals, the intact adult male rat assay has the capacity to detect estrogen receptor agonists/antagonists, androgen receptor agonists/antagonists, progesterone receptor agonists/antagonists, steroid biosynthesis inhibitors, gonadotropin and thyroid modulators either directly or indirectly by altering the hypothalamic-pituitary-gonadal or -thyroidal axes, and prolactin modulators through neuroendocrine pathways.

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 or negative effect on the estrogen, androgen or thyroid hormonal systems.

The purpose of this peer review is to review and comment on the intact adult male screening assay for use within the EDSP to detect various MOAs, especially AR agonists/antagonists, steroid biosynthesis inhibitors, gonadotropin and thyroid modulators either directly or indirectly through intact HPG or HPT axes.

Although peer review of the intact adult male assay will be done on an individual basis (i.e., its strengths and limitations evaluated as a stand alone assay), it is noted that this assay along with a number of other *in vitro* and *in vivo* assays will potentially constitute a battery of complementary screening assays. A weight-of–evidence approach is also expected to be used among assays within the Tier-1 battery to determine whether a chemical substance has a positive or negative effect on the estrogen, androgen or

thyroid hormonal systems. Peer review of the EPA's recommendations for the Tier-1 battery will be done at a later date by the FIFRA Scientific Advisory Panel (SAP).

CHARGE QUESTIONS

Each peer reviewer is asked to review the Integrated Summary Report and accompanying support materials and comment on the results of the validation process of the 15-day intact adult male rat assay, especially the inter-laboratory validation exercise.

1. Please comment on the clarity of the stated purpose of the assay.

2. Please comment on the clarity, comprehensiveness and consistency of the data interpretation with the stated purpose of the assay.

3. Please comment on the biological and toxicological relevance of the assay as related to its stated purpose.

4. Please 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.

5. Please comment on the strengths and/or limitations of the assay.

6. Please 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.

8. Please provide comments on repeatability and reproducibility of the results obtained with the assay, considering the variability inherent in the biological and chemical test methods.

Appendix B

INTEGRATED SUMMARY REPORT

Integrated Sumary Report for Validation of 15-Day Intact Adult Male Rat Assay as a Potential Screen in the Endocrine Disruptor Screening Program Tier-1 Battery (PDF) (223 pp, 2.7M)

Appendix C

SUPPORTING MATERIALS

The following supporting documents were provided to reviewers:

- Final Report - Inter-Laboratory Validation of the 15-Day Adult Intact Male Rat Assay with Linuron and Phenobarbital (Charles River Laboratories) (PDF) (458 pp, 17.3M)
- Final Report - Inter-Laboratory Validation of the 15-Day Adult Intact Male Rat Assay with Linuron and Phenobarbital (RTI International) (PDF) (479 pp, 16M)
- Final Report - Inter-Laboratory Validation of the 15-Day Adult Intact Male Rat Assay with Linuron and Phenobarbital (WIL Research Laboratories, LLC) (PDF) (643 pp, 15M)