Peer Review Results for the Male Pubertal 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

01 November 2007

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

TABLE OF CONTENTS

Page

LIST OF FIGURES (Continued)

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 male pubertal rat assay. The purpose of the pubertal assay is to provide information obtained from an in vivo mammalian system that will be useful in assessing the potential of a chemical substance or mixture to interact with the male endocrine system. This assay is capable of detecting chemicals with antithyroid, androgenic, or antiandrogenic [androgen receptor (AR) or steroid-enzyme-mediated] activity or agents which alter pubertal development via changes in gonadotropins, prolactin, or hypothalamic function.

Although peer review of the male pubertal rat 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 expected to be

used among assays within the Tier 1 battery to determine whether a chemical substance interacts with the endocrine system. Peer review of the EPA's recommendations for the Tier 1 battery will be done at a later date by the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) Scientific Advisory Panel (SAP).

The purpose of this peer review was to review and comment on the male pubertal rat screening assay for use within the EDSP to detect chemicals with antithyroid, androgenic, or antiandrogenic [androgen receptor (AR) or steroid-enzyme-mediated] activity or agents which alter pubertal development via changes in gonadotropins, prolactin, or hypothalamic function. 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 pubertal 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 male pubertal rat assay on September 17, 2007. ERG held a pre-briefing conference call on October 3, 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 October 22, 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 male pubertal 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 somewhat arbitrary as in some instances it is equivocal as to which of the questions posed they address. I have little expertise with regard to the thyroid axis, so have restricted my comments on this aspect, and evaluation by those with proven expertise on this axis should be heeded rather than my own comments in this regard.

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)]*. One of the test systems recommended by the EDSTAC was the male pubertal rat assay. The purpose of the pubertal assay is to provide information obtained from an *in vivo* mammalian system that will be useful in assessing the potential of a chemical substance or mixture to interact with the male endocrine system. This assay is capable of detecting chemicals with antithyroid, androgenic, or antiandrogenic [androgen receptor (AR) or steroid-enzyme-mediated] activity or agents which alter pubertal development via changes in gonadotropins, prolactin, or hypothalamic function. 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 a well-written protocol and a well-written and well-managed validation study. Clearly, the EPA thought about performance criteria and about the logic required to interpret the findings. The performance criteria should be sharpened, both or the way the endpoints at necropsy are evaluated as the RIA performance criteria. There are additional comments, aside from the responses to the charge questions, that will be reviewed first.

General Comments:

Page 9, line 3, "…but these were later removed from the protocol as being relatively uninformative due to wide variation in levels."

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.

Page 24, line 7, "Although the primary cellular mechanism for this compound's effects on endocrine function are not characterized, it is well established that atrazine disrupts the hypothalamic (central nervous system, CNS) control of pituitary function by suppressing the gonadotropin releasing hormone (GnRH) stimulation". This statement does not appear to be referenced. Certainly, the work by Cooper et al., (1999 and 2000), and the work of McMullin et al. (2003) cannot be used to support this statement.

Page 25, lines 13-24, "…The conclusion from this study was that the pubertal male assay clearly identified atrazine as interacting with the endocrine system at both dose levels, thus showing that the assay is sensitive to chemicals that affect the HPG axis….". EPA's conclusion is dangerous! Rather, this study shows that when the EPA has previous information indicating that a chemical interacts with the endocrine system, they can selectively interpret the data in a way that is consistent with what is known. It is important to recognize that this entire paragraph amounts to arm waving. It would be interesting to see what would happen if the EPA were to test chemicals in a blinded study in which neither the contract lab, nor the EPA interpreters were aware of the test chemical.

Page 26, lines 9-14. This conclusion ignores the observation made directly above it that serum T_4 levels were reduced, although serum TSH was not altered, nor were there treatment-related histological changes in the thyroid gland. No basis is given for ignoring T_4 levels.

Page 28, lines 8-12. This is a rational statement, but the logic is not spelled out.

Page 28, lines 14-18. This paragraph illustrates a weakness in the EPA's logic. First, low thyroid hormone can cause a decrease in weight gain; thus, animals treated with high levels of PTU could have a lower body weight precisely because serum T_4 levels are lower. In addition, serum T_4 levels are sensitive to caloric restriction; therefore, if animals treated with a high dose of compound such that caloric intake is restricted, serum T_4 levels could be lower due to this mechanism.

Page 31, lines 20-22. "Thyroid weight was increased at both doses, though the increases were not statistically significant (27.3 mg in controls, 31.9 and 32.5 mg at the low and high dose levels, respectively). As a fundamental rule, if something is not statistically significantly different, it is not different. Also as a fundamental rule, the biological significance of a difference may be arguable, but if an endpoint is statistically significantly different, it should be reported and interpreted as such; and if not….

Page 32, Lines 27-31. "The conclusion of this part of the study was that the male pubertal protocol appears to be sensitive to the thyroid-related and gonadal effects of Phenobarbital even though the thyroid-related responses were not significant at the $p<0.05$ level." This exact same profile was observed with Linuron and with Flutamide, yet the EPA concluded that the thyroid

endpoints were not important. ***These studies are showing only that EPA can identify a well-known endocrine disruptor, not that they can identify an endocrine disruptor for which there are no previous data***.

Page 34, lines 10-11. "Due to an oversight, serum hormone levels (T_4, TSH, testosterone) were not obtained in this study." This demonstrates that "GLP" means only that record keeping is precise, not that the study was performed according to plan, or that the techniques used to perform the study were appropriate or adequate.

Page 39, lines 5-7. "…a chlorotriazine herbicide which had recently…." This statement lacks foundation. The paper by Stoker et al. 2000a does not show this. My search of MEDLINE did not reveal studies that would support this statement. My query of investigators in this field also did not reveal studies that would support this statement.

Page 40, line 10-12. "At this point, no environmental chemicals have been found to bind to the thyroid receptor. (See Stoker et al. (2000b) for review of toxicant effects on thyroid function.)" **This statement is incorrect. Several compounds have now been shown to bind to the TR, some with IC50's in the nM range. The authors are not even using the EPA-managed thyroid DRP.**

Page 50, Table 17. It is not clear how these data were derived.

Page 51, Table 17. The performance standards for the RIAs does not take into consideration that the contract labs are reporting performance of their assay that falls outside the performance standards reported by the manufacturer.

2.2 Comment 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. 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. 5, line 7). 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 line 8, p. 5. 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 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. The report does a good job of describing the selection of test chemicals for validation as being based on their mode of action. It should also be made explicit that successful validation would include results from the assay that correctly identify those modes of action.

The pubertal male assay is a complicated assay with a large number of endpoints. The purpose of the assay is also complex, with the intent to detect a large number of modes of endocrine action. The basic concept for the assay is that agents with a specific mode of action would affect only a subset of the endpoints measured, and that by analyzing which endpoints were affected it

would be possible to discern mode of action. It would be very helpful to the reader to have a table early on in the report that describes which endpoints are expected to be affected for each mode of action being evaluated. This information is contained in the descriptions of the effects of the test agents used in the various studies, but it would be much easier to interpret the results if a summary table linking endpoints with mechanisms were provided.

Richard Dickerson: The first paragraph of the stated purpose of the assay is clear enough but could be improved by eliminating or replacing some of the phrases. For example, the phrase "information that will be useful in assessing the potential of a chemical substance or mixture to interact with the endocrine system" is much too long and passive. Consider replacing it with "information useful in determining the potential of chemicals or mixtures to interact with the endocrine system.

In the second paragraph, it is stated that the test compounds will be dissolved in corn oil whereas test substances are dissolved in a methyl cellulose vehicle for the 15-day adult male assay. Why the difference in choice of vehicles? Corn oil contains varying amounts of phytoestrogens that may mask some changes in the endocrine system. It seems like a more inert vehicle such as triolean would be more appropriate. In addition, pups should be culled to 8 or 10 per litter rather than the range of 8-10 and housed 2 or 3 per cage instead of 2-3 to further reduce variability. There is no strain of rat listed in this brief description of the assay; shouldn't it be stated here?

Kevin Gaido: The stated purpose of the male pubertal assay, is to provide an alternative to the female pubertal assay as an in vivo mammalian system useful for detecting chemicals that interfere with androgen or thyroid function, or alter hypothalamic function, gonadotropin or prolactin secretion. The assay will ultimately become part of a comprehensive battery of tests for endocrine active chemicals.

Richard Sharpe: The background information and protocol description give a clear view of the objectives of the assay and the role of its component parts. This material should be easily comprehensible to anyone intending to use and apply this assay. It is a Tier-1 assay and its priority is to maximize the detection of endocrine active compounds (that target the male reproductive and/or thyroid axes) whilst minimizing false negatives. The use of multiple

endpoints (mainly organ weights) is designed to ensure this. The use of an intact animal with normally functioning, homeostatic, hormone systems is, in my opinion, a strength as it represents the 'real world'. It includes metabolism etc and can potentially 'integrate' chemical effects that may directly affect one hormonal (or non-hormonal) system and then secondarily affect another hormonal system.

Tom Zoeller: The purpose of the assay is clear. It is difficult to imagine what a novice in this field would require to perform the assay as intended by the EPA; presumable, the contract labs performing this would have experience.

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

George Daston: The report describes in the context of the validation framework a large number of studies that were conducted for a variety of purposes. These include transferability of the assay, reliability, and validity of the assay for detecting several modes of action. For the most part, I found the presentation to be clear and the interpretation consistent with the results of the studies. For the purpose of clarity in this review, I will address interpretation issues for each of the study types separately, and then the overall interpretation of the studies in aggregate.

Transferability study: This study was the first conducted outside of an investigative research lab and confirmed that the protocol yielded similar results in a separate laboratory. The study was extensive and evaluated six chemicals representing different modes of action, in two strains of rats. The results are thoroughly presented, and I agree with the overall interpretation that the protocol is transferable. That said, the claim of transferability would be better supported if data from similar studies (either from EPA's labs or from the literature citations described for each chemical) were presented along with the contract lab results. Transferability involves getting the same response (at least qualitatively) in two different labs. The text of the document indicates that this was the case, but it would help the reader to understand this if the original data were summarized in tabular form for easier comparison with the Therimmune data.

Page 12, lines 18-19, indicates that the research at Therimmune evaluated the transferability of the protocol "as it existed in 1999". It would be useful to describe briefly what the differences

were so that the reader can be convinced that the conclusion of transferability is relevant to the current protocol. I assume that the difference was limited to a few endpoints that were not measured at the time (thyroid weight, testosterone) but that was not clear from reading the report.

The transferability study is also noteworthy in that it adequately addresses the question of strain sensitivity and supports the decision to use Sprague Dawley rats in subsequent studies.

Assay Sensitivity: This section describes studies to evaluate many more chemicals representing a range of modes of action, to evaluate multiple dose levels, and to assess the extent to which changes in body weight gain affect the other endpoints measured in the study. Again, the results are thoroughly described. I agree with most of the interpretation, but there are a few conclusions that are troublesome and that I believe should be looked at again.

The conclusion about phenobarbital (p. 32, lines 27-29) is that the protocol was sensitive to the thyroid-related effects of Phenobarbital despite the fact that there were no statistically significant changes in most of the thyroid-related responses. The report blames the lack of significance on the fact that the MTD wasn't reached and that therefore the experiment was not an adequate test of the resolving power of the assay. I can accept that rationale, but if true, then nothing can be said about the ability of the assay to detect Phenobarbital as an indirect thyroid toxicant. The preceding text on p. 32 is clear about that point, but it should be reflected in the concluding paragraph.

The second result that deserves some additional discussion is the number of compounds that affect thyroid hormone (T4) concentrations. It appears that T4 is not a particularly specific measure of thyroid toxicity. This lack of specificity has been noted in other screening protocols (e.g., O'Connor et al., 2002) and should be discussed in this report in some depth, either in this section or in the overall conclusions. It is possible that T4 measurements can only be interpreted in the context of some other effect to be meaningful. The more that this is discussed before this or other assays come on line, the better the interpretation of results will be for compounds with uncharacterized modes of action.

The multi-dose study is thoroughly described. There is only one point on which I believe the interpretation should be expanded: p. 36, lines 4-7, it is explained that an apparent result of flutamide on adrenal weight was probably due to inordinately low values in two control animals and was not an effect of the compound. I agree with this conclusion, but it raises the question of whether this apparent variability problem is important enough to correct. For example, in another part of the report, it was determined that high variability in the weights of fluid-filled organs indicated that more detail and training was needed across labs in proper dissection. In this instance, could it also be a problem with procedure, or is the variability due to the animals themselves, in which case it may be important to either increase Ns or to amass a larger historical control data set against which new results can be compared.

The section describing the study on the effects of body weight gain is satisfactorily described. There is one semantic correction that I would like to see made in this section, which is to change all references to "body weight loss" to "decreased body weight gain". The latter is actually what was measured. The animals on restricted feed were not losing weight; they were gaining weight at a slower pace than controls. This distinction is important because a misunderstanding of this point could lead to inappropriate dose selection, excessive toxicity, and an uninterpretable study.

The final paragraph of this section (p. 47-48) needs to be clarified or expanded on. The first sentence of the paragraph states that a 10% reduction in body weight is a reasonable basis for setting an upper limit in the assay. However, the next sentence suggests that this level of effect is too severe for thyroid effects and the final sentence indicates that a 9-10% decrease may necessitate conducting additional studies and/or a weight-of-evidence approach to interpret the results for the thyroid endpoints. This seems to indicate that the male pubertal assay cannot be optimized for thyroid endpoints and reproductive endpoints, and that either a compromise would need to be made in selecting an upper dose level, or an additional assay would have to be run. This is an important point in determining whether the assay is a suitable alternative, and should be discussed at greater length.

Performance Criteria: The section on performance criteria is adequately presented, but Table 17 needs some attention. The column heading "1.5 CV" must be incorrect (since the values are less than those for CV in the column immediately to the left). I believe it should be "1.5 SD". Also,

the last column would be more understandable if it were titled something like “Maximum acceptable CV”.

Interlaboratory study to examine reproducibility: This section adequately describes the study and its results. There are only a few points that I found troubling. First, on p. 52, lines 14-18, it is stated that EPA does not require the assay to consistently display a pattern of endpoint responses diagnostic for a particular mode of action, but only that thyroid responses not be used to claim consistency with sex steroid associated responses and vice versa. Given that the stated purpose of the assay is to detect a variety of hormone-related modes of action, and that, as the most apical assay in the screening battery, it will have the greatest influence on weight of evidence determination of the battery, we need to expect more of the assay.

The final paragraph on p. 72 indicates that, while the assay results are generally reproducible, the CV criteria that were established a priori were not always met. I don’t see this as a fatal flaw with the assay, but it will be important in the early going to constantly re-evaluate the magnitude and sources of variability and to find ways to minimize the latter and better set acceptable criteria for the former.

Richard Dickerson: The data from each of the laboratories was presented clearly and factually. Several areas of concern were discussed including strain variability, variability in determining the day of PPS, and measurement of weight of fluid-filled organs. However, reasonable solutions were proposed for each of these areas of concern.

For the multiple chemical studies, the effects of phenobarbital exposure on the thyroid axis were noted as non-significant but the trends were in the correct direction. One reason suggested for the lack of significance was failure to reach the MTD. Failure to reach the MTD may be common when there is a lack of data concerning the toxicity of test compounds. Perhaps a quantitative method of determining significance when a number of endpoints almost reach the stated level of significance could be used or the standard could be multiple endpoints reaching the 0.10 level of significance as a way of determining weight of evidence.

Kevin Gaido: The summary statement provides a clear and comprehensive interpretation of the available data. A comprehensive review of the literature is presented and considered. A detailed comparison of the results from each laboratory together with historical data is provided. Results from the interlaboratory validation demonstrate that the protocol is transferable and reproducible and capable of detecting chemicals that act through a variety of endocrine related mechanisms to impact male pubertal development.

Richard Sharpe: A considerable amount of data has been accrued on this assay and has involved several different (but experienced) laboratories and the testing of a large number of compounds with a wide variety of purported or known mechanisms of action (MOA). The evidence presented for review and in a few publications substantiate the view that this assay is fit for purpose. As with the intact adult rat assay, the pubertal rat assay uses multiple endpoints in order to maximize detection of 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). It is simpler than the adult male assay in terms of endpoints (less hormonal data), and arguably this makes the assay potentially more straightforward to operate and interpret, though it requires a considerably longer treatment period. The use of multiple endpoints may provide preliminary information on the potential MOA but, 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 several inter-laboratory comparisons, are rational and fit with current understanding of how the various endocrine systems operate within the body during pubertal development. There are some issues in relation to assay specificity and which endpoints are absolutely essential, and others which might be dispensable or assigned a ‘supporting role’ (see below), but these are rather minor issues, and they do not affect the overall conclusion that the assay appears robust and fit for purpose, but with some limitations.

Although results obtained in applying the assay to a variety of chemicals with known endocrine activity have been largely as expected, the specificity of the assay is to some extent unproven. In

the inter-laboratory study, 2-chloro nitrobenzene (2-CNB) was included as a test of specificity as various indicators had suggested that it would not have 'endocrine activity' and would thus provide a test of assay specificity. However, 2-CNB significantly reduced testosterone levels, weights of all androgen-dependent organs and delayed preputial separation as well as causing histopathological changes to the testes. Based on these findings, 2-CNB would clearly be classed as an endocrine disruptor and it must be presumed that this is the case (nitrobenzene and dinitrobenzene are well-established testicular toxicants in the adult rat, although I am not aware of evidence that they disrupt Leydig cell function). The alternative interpretation is that the pubertal assay is prone to non-specific effects and will therefore yield 'false positives' with a high frequency. Although it is preferable for a Tier-1 (screening) assay to suffer from this rather than from frequent false negatives, it will be important in future studies to more rigorously investigate the specificity of the pubertal assay. My expectation is that the assay will not be unduly prone to false positives, based on the relative lack of impact of the feed restriction studies on the assay endpoints, but this needs to be demonstrated categorically using appropriate test compounds.

Tom Zoeller: A significant weakness is that the interpretation of the thyroid endpoints seems to require previous knowledge of the activity of the test compound. The justification of this statement is that the profile of effects observed with Phenobarbital was identical to that observed with other compounds like Linuron and Flutamide, yet the interpretation was based on previous publications. Moreover, all hormone levels are not being measured in a manner that will lead to logical interpretation.

2.4 Comment 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 IV (p. 9). I believe that this assay, as well as the adult male and pubertal female assays being evaluated, has the potential to provide the most reliable and comprehensive information for the weight-of-evidence determination. 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 pubertal male also has special relevance in that puberty represents a major developmental stage in the maturation of the reproductive system, and may be particularly susceptible to exogenous agents that interfere with the hormonal control of sexual maturation. Because young, rapidly growing animals are used, the system is expected to be sensitive to agents that affect thyroid function. Stoker et al. (2000) provides an excellent review of the literature on the effects of exogenous agents on puberty and thyroid function in the pubertal rodent. This review provides a strong theoretical underpinning of the relevance of the assay for the stated purpose of screening for potential effects on androgen and thyroid-dependent systems, and possibly other modes of action.

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. These endpoints are sensitive to exposure of known androgen and thyroid agonists and antagonists. The endpoints used for the HPT axis are also the most appropriate for the length of the assay. Therefore, the assay measures physiological endpoints appropriate for detecting alterations in the status of the male reproductive and thyroid organ systems.

In terms of toxicologic relevance, the endpoints selected for the Pubertal Male Rat Assay are appropriate for several reasons. First, they reflect biologically relevant endpoints as discussed above. Second, validation 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. Third, exposure to a dopamine antagonist, pimozide, showed that the assay was sensitive to compounds that inhibit prolactin

release. Fourth, exposure of test animals to propylthiouracil, a thyroid hormone synthesis inhibitor, and to phenobarbital, which increases metabolism of thyroid hormones, demonstrated the assay can detect compounds that alter the production or clearance of thyroid hormones. 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: The assay was designed to detect chemicals that interfere with androgen or thyroid function or with the HPG axis based on the understanding of the biological relevance of these functions for normal pubertal development. Serum hormones and reproductive organ weights significantly increase in male rats during puberty and as a result, chemicals that disrupt endocrine function can have a dramatic impact on male pubertal developmental measurements such as organ weights and preputial separation. This assay is highly relevant for toxicological screening for endocrine active chemicals.

Richard Sharpe: The validation process, which involved testing of a considerable number of compounds with a range of known activities and mechanisms of action provided solid foundations for subsequent evaluation and interpretation of results for compounds of unknown hormone activity. It also identified unexpected effects or results, such as those for 2-CNB and phenobarbital, which are discussed elsewhere. It is likely that continued application of the assay to a wider range of chemicals will uncover other activity profiles that do not fit within our expected concepts, but this will inevitably lead to a better understanding of the utility, and limitations, of the assay. I anticipate a progressive ability to categorize chemicals into classes based on their activity profile in this assay, even if it is not possible to define a clear MOA. Because the test uses an intact animal that is advancing through one of the most endocrinologically dynamic phase of its life (puberty), it might be anticipated that compounds might affect target organs or hormone levels via pathways that are unrelated to endocrine disruption *per se*, for example effects on food intake/metabolism that lead secondarily to such changes. However, from results obtained so far, and including food restriction studies, this expectation has not surfaced in any major way, which arguably makes the assay more robust, simpler in its function and analysis and relatively easy to interpret.

Tom Zoeller: The assay is relevant to the goals of the EDSP. Data from this assay will not likely provide novel biological information, although it could provide the motivation to address specific mechanistic hypotheses.

2.5 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

Richard Dickerson: My comments are based on the protocol attached as Appendix 1.

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. Methods for housing and treatment of the animals are presented in sufficient detail. Each endpoint is clearly described and methods for statistical analysis as well as how to handle outliers are presented. Finally details and examples for data interpretations, presentation, and developing a final report are given. The reproducibility and transferability of the assay is clearly demonstrated by the reproducibility of overall results across laboratories. While there was some variability with some endpoints between the laboratories the overall weight of evidence and conclusions were consistent.

Richard Sharpe: Insofar as I feel competent to judge (as a scientist running an academic research laboratory), the protocol provided is clearly laid out, is understandable and is sufficiently detailed to enable an appropriately experienced laboratory to run, complete, evaluate and report results using this assay. There are no deficits in the protocol that I have noticed. The fact that only 3 hormonal assays are included, namely for thyroxine (T_4), thyroid stimulating hormone (TSH) and testosterone, makes the pubertal assay relatively easy to run, as hormone assays are often a source of considerable inter-laboratory variation and, because they provide only a 'snap-shot' in time of hormone levels, they can be potentially misleading.

Arguably one of the more variable and more subjective aspects of the pubertal assay is the endpoint of preputial separation (PPS). This is clearly a useful 'endocrine' endpoint that summates androgen action over a period of time, as shown in all of the studies so far done to test

and hone the pubertal assay. Nevertheless, in the inter-laboratory study none of the 4 (experienced) labs involved could meet the coefficient of variation (CV) criterion for bodyweight at PPS and only 2 could do so for age at PPS. Some refinement of the PPS 'definition' was adopted after issues relating to incomplete PPS (retention of 'threads' of connecting tissue), but by the very nature of the assessment it seems to me that PPS will always be prone to high between-laboratory and between-observer variation. Similar to PPS, the recorded weights of the ventral prostate, seminal vesicles + coagulating gland, epididymis and levator ani +bulbocavernosus muscle all provide a measure of androgen action over time – in essence they summate androgen action over the 30-day course of the assay. Accordingly, weights of these organs changed more or less in parallel to PPS in response to exposure to the various compounds tested in the different parts of the validation exercise and in other studies. My initial reaction to this (expected) observation was that PPS was maybe redundant, and could therefore be dispensed with, as it did not measure anything that the target organs already did not. However, it is also apparent that for most of these organ weights there is similarly high CV as discerned from the inter-laboratory study. Thus, none out of the 4 labs met the CV criterion for ventral prostate weight, only 1 met the CV for seminal vesicle weight and only 2 did so for epididymal weight. None of this is unduly surprising, as anyone experienced in the dissection of these organs will know that not only is there high variability in actual weight (probably largely reflecting different levels of fluid content), but the dissection process can also be variable depending on how this is done. For this reason it is good practice to have the same person do all of the dissections for the same organ in order that variation within a laboratory is minimized.

Returning to the issue of whether PPS is worth retaining, I convinced myself that it was, based on two lines of reasoning. First, it is an 'in-life' measurement, and thus may provide the first indication of 'anti-androgenic' or 'androgenic' activity in a test compound which can then be confirmed by organ weight measurements. Second, as PPS and reproductive organ weights are all intrinsically highly variable measurements (for reasons outlined above), it is safer to have multiple endpoints that reflect the same underlying phenomenon/activity (ie. androgen action over time), as this will increase the chances of detecting a significant effect on any one of the endpoints; the fact that one is in-life and the others terminal reinforces this argument. Additional to this reasoning is that PPS is non-invasive and not time-consuming as the visual inspection can be made at the same time as dosing the animal.

2.5.1 Comprehend the Objective

George Daston: This section is succinct and to the point.

Richard Dickerson: Section I. Purpose and Applicability is clearly worded and states the purpose of the assay precisely in such a way that individuals with scientific training can easily comprehend it. It might require re-phrasing for a less technically audience or addition of a lay summary.

Tom Zoeller: The objectives are clearly and concisely articulated in the protocol.

2.5.2 Conduct the Assay

George Daston: The protocol contains enough information for a competent laboratory to conduct the assay in a consistent way. There are a few aspects of the protocol that may be too restrictive, such as the admonition to keep temperature and relative humidity within ranges that may not be achievable in all facilities, and are different from those established by AAALAC.

Richard Dickerson: The instructions on how to conduct the assay are complete and clear for the most part. Prohibition of the use of tap water for the animals is very good to eliminate perchlorates and other contaminants. Specifying rat chow that is low in phytoestrogens is good but consider adding as an appendix a list of acceptable rodent chows. Replace the words kill, killed or killing with euthanized or euthanasia. The humidity conditions are perhaps too stringent and may prevent the assay from being performed in regions of the country that are too dry or too humid. For example, facilities in West Texas, New Mexico and Arizona frequently can not maintain 30% humidity. Section III states that the litters will be standardized or culled to 8-10 pups per litter. For a particular study, all litters should be standardized to the same size. Why is it not acceptable to cross-foster to raise the litter size to the minimum? It can prevent waste of animals if used properly. Section VI discusses the properties of corn oil to be used but does so in qualitative terms. Consider changing it to corn oil from an approved source, from a freshly opened container, and free of sediment. However, the use of corn oil is problematic in that it can contain phytoestrogens. Why not use an synthetic oil such as triolean? Section IX

needs to be revised to meet AVMA Panel on Euthanasia standards. See the note below. The remainder is fine.

Tom Zoeller: To the best of my ability to determine, the protocol is sufficiently detailed that an experienced laboratory could conduct the assay as written.

2.5.3 Observe and Measure Prescribed Endpoints

George Daston: The information in the protocol and attachments were clear and helpful in providing guidance on evaluating the endpoints. I found it very useful that the attachments to the protocol included information useful to conducting the assay, such as reference images for thyroid histology.

Richard Dickerson: clear and concise

Tom Zoeller: Generally, the prescribed endpoints are clearly articulated. In addition, the EPA has developed performance standards, which will likely improve the quality of the data received from this assay. However, the mechanism by which these performance standards are generated should be more closely evaluated. Moreover, there are no performance standards established for the RIAs. These commercial RIA kits come with manufacturer-established performance characteristics, but the EPA does not require that contract labs use these kits in a manner that is consistent with the manufacturer performance. Finally, some of the kits being used in this assay are heterologous (i.e., prepared and calibrated for human samples, but used in rats), and the EPA does not require the contract lab to validate the assay. Given this situation, it is no wonder that there is a high degree of variation in hormone measurements. It is highly likely that the variability in hormone levels observed in these experiments can be reduced to such an extent that hormone levels themselves can play a larger role in the in vivo portion of the EDSP.

2.5.4 Compile and Prepare Data for Statistical Analyses

George Daston: Sufficient guidance was provided on how the data should be displayed and analyzed. The statistical procedures were not overly restrictive but provided enough guidance to facilitate comparison of study results in different labs.

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

Tom Zoeller: The protocol is clear in directing laboratories in their data preparation and analysis.

2.5.5 Report Results

George Daston: The protocol provided a great deal of information on how to report and interpret results. The length of the data interpretation section is unusual in my experience, but given the novelty of the protocol and the complexity of interpretation, I believe it to be warranted.

Richard Dickerson: clear and concise.
Notes:

1. Decapitation without prior use of either CO2 or inhalational anesthetic is not an AVMA Panel on Euthanasia (2000) approved method of euthanasia. It is listed as a conditionally acceptable method by the Panel but requires justification for use such as peer-reviewed journal articles that conclude prior exposure to CO2 or inhalational anesthetic alters the endpoints that are the focus of the study. The statement is made in the protocol that decapitation is considered more humane than CO2 asphyxiation. This is in direct opposition to the AVMA Panel on Euthanasia which is reference used by IACUCs, AAALAC, the NIH Guide, and the USDA.

2. Heat-treated aspen shavings are preferred to heat-treated pine shavings since they contain lower levels of volatile compounds and appear to be less allergenic to workers.

Tom Zoeller: Likewise, the kind of information requested in the report from studies are clearly represented in the protocol.

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

George Daston: The section on pp. 90-93 of the report does a good job of summarizing the strengths and limitations of the assay. I consider the major strengths of the assay to be:

a. The assay is performed in an intact system, capable of responding to interactions among multiple endocrine axes.
b. The assay uses developing animals, which may be more susceptible to endocrine perturbations, thereby potentially increasing the sensitivity of the assay.
c. The assay measures endpoints indicative of a number of modes of action of interest to EPA for endocrine screening. It should be possible to obtain a lot of information out of a relatively few animals.
d. The apical nature of the assay will allow it to provide a lot of context for weight-of-evidence interpretation of some of the simpler assays in the proposed tier 1 battery. It also allows for multiple endpoints for each mode of action, improving the interpretability of the assay.

The limitations of the assay are

a. The assay is lengthy. Dosing covers a 30 day period. Furthermore, the study requires a pilot study for dose-setting. The length of the assay is not conducive for rapid screening.
b. The assay appears to be inordinately influenced by changes in body weight gain, such that a significant change in body weight gain is needed for a valid (presumably, sensitive) assay, but changes above 9-10% make the assay difficult to interpret. It may not be possible to achieve this much precision in dose setting, especially in a screening context with limited numbers of animals.
c. There is not enough information to assess the specificity of the assay. It may be too non-specific to provide enough information to correctly classify negatives. More research needs to be done to address this concern.

d. There is not enough information yet to make definitive conclusions about the range of modes of action that are detectable by the assay. The evidence is good that the assay can detect anti-androgens and thyroid-active agents. There is also reasonable evidence that the assay can detect androgens, agents that affect the hypothalamic-pituitary gonadal axis, and steroidogenesis inhibitors. The evidence on estrogens and aromatase inhibitors is mixed. These may not be limitations of the assay, but are limitations on how we can interpret the utility of the assay at present. Clarity on which modes of action it covers will be important in determining the assays that should be conducted with it as part of a battery.
e. The apparent non-specificity of T4 is concerning, and may suggest that this measure cannot be interpreted out of context with other thyroid measures

Richard Dickerson: Strengths of the assay include ease of conducting the assay and measuring the endpoints, the moderate duration of exposure, biological and toxicological relevance and substantial literature base. In addition, this assay uses intact animals whereas other assays such as the Hershberger assay require castrated animals. Furthermore, many of the techniques used are common to other assays such as the 15-day adult rat assay, multigenerational testing protocols, and Hershberger assays that have been used for some time. The use of time of PPS as an endpoint is an advantage due to its increased sensitivity to androgen agonists and antagonists compared to weight of androgen-responsive organs and tissues. Coupling this assay with the pubertal female rat assay provides a robust test capable of detecting xenoestrogens, antiestrogens, androgen agonists and antagonists as well as inhibitors of thyroid synthesis, inducers of thyroid hormone metabolism and compounds that alter release of pituitary hormones.

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. This assay does not differentiate between the various mechanisms of action by which a compound can affect androgen status or whether change in thyroxine is due to inhibition of synthesis or induction of metabolism. This assay also uses timed pregnant rats which are not only live animals but fairly expensive. Although the exposure period is short in comparison to some other assays, it does

require 30 days. Last, this assay is more recent than the Hershberger assay or the rat uterotrophic assays which translates into a smaller data base and less harmonization.

Kevin Gaido: Strengths of this assay include the ability to screen for multiple modes of action in a sensitive *in vivo* mammalian assay. The assay has extensive historical data from multiple laboratories and the biology behind the various endpoints is well understood. The assay focuses on a time period when reproductive organ development is very sensitive to endocrine disruption. 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. The assay has multiple and redundant sensitive endpoints that can be used to help design more definitive Tier-2 testing. For example, results with 2-CNB provided sufficient information to suggest that its effect on the growth of androgen dependent tissues is either through altering steroidogenesis or targeting the secretion of pituitary hormones but not through interference with the androgen receptor.

A weakness of this assay may be its ability to screen for week estrogens. While the assay can detect more potent estrogens, there is insufficient data regarding weaker estrogens. In addition, the lack of a test of a negative control makes it difficult to determine the specificity of this assay for endocrine active compounds. The reliance on MTD is a weakness since additional prior studies must be performed to accurately identify MTD or in cases where the MTD is based on a review of the literature may lead to over or underestimating the MTD.

Richard Sharpe: The four main strengths of the assay are: (1) that it has a strong foundation based on various studies that have been undertaken as part of the validation exercise using a variety of compounds with different activities; (2) the primary reliance on target organ weights, which are easily measured, as the most predictive endpoints; (3) the use of multiple endpoints to inform on the same 'hormonal activity', some of which extend beyond organ weights (histopathology, hormone levels) and one of which is a dynamic pubertal 'in-life' event measured over time (PPS); (4) the minimal use of hormone measurements, as these are not only more technically difficult to measure, but may be difficult to interpret at a time in development when there can be large fluctuations within, and especially between, animals.

The main weakness of the assay, which at this stage of evaluation is more theoretical than experience-based, is that exposure of rats to compounds during such a hugely dynamic phase of growth and reproductive development as puberty is likely to result occasionally in effects on reproductive 'endocrine systems' that are not due to intrinsic 'endocrine activity' but to an indirect effect eg. on growth. Although the feed restriction studies did not produce evidence to suggest major impact of this on reproductive organ weights or PPS in the pubertal assay, I still consider this to be a likely event with some compounds; the effects of phenobarbital are worth discussing in this respect (see below). Nevertheless, such an outcome will mean at the worst that some compounds are identified as 'false positives' which is acceptable in any screening assay provided this is not unduly frequent; the issue of assay specificity, which remains to be resolved, is important in this regard and has been discussed above.

I was somewhat disconcerted to see that, in both the multi-chemical study and in the TherImmune2 study, phenobarbital was detected as a clear anti-androgen, based on delayed PPS and reduced reproductive organ weights; equally disconcerting was its lack of significant effect on the thyroid axis. My presumption was that phenobarbital had been included as a test compound primarily because of its potential thyroidal effects (clearly evident in the adult male rat assay) and because its central effects might also 'spill over' into effects on the reproductive axis. As there is published data to show that phenobarbital can suppress LH secretion and even inhibit (fetal testis) steroidogenesis, the effects in the pubertal assay are not entirely unexpected but I was surprised at their robustness. It is possible that the pubertal rat is especially vulnerable to central effects of phenobarbital as normal progression of puberty is dependent on a progressive increase in frequency and amplitude of LH pulses, which then drive increasing testosterone production; the alternative interpretation, namely that the pubertal rat is more susceptible to non-specific effects remains to be resolved, as discussed above.

Tom Zoeller: The strength of the assay is that it evaluates the interaction of toxicants with the androgen and thyroid systems during a particularly sensitive period of development for these interactions. Thus, this assay should be more sensitive that the adult in identifying EDCs. The endpoints for androgen action are valid and known to be sensitive to changes in hormone action during this period. There are no endpoints of thyroid hormone action that are equivalent to those of the androgen system (e.g., secondary sex organ weight). This produces an unbalanced assay

that has confounded the interpretation of results of this assay and the Adult Intact Male 15-day Assay. Specifically, several compounds cause a reduction in serum total T_4, including Linuron, PBDEs, PCBs, Phenobarbital. In each case, the interpretation is that these compounds activate liver enzymes (e.g., UDPGTs) that decrease circulating half-life for serum T_4. However, in the case of Phenobarbital (mostly), serum TSH is increased in response to low T_4. In contrast, serum TSH is not always elevated in response to low T_4 produced by Linuron, or PCBs. It is not clear why the same level of serum T_4 is not always associated with an increase in T_4, but this apparently is the case. The question is whether low serum T_4 produces adverse effects on peripheral tissues – especially during development – in the absence of increased TSH. Failure to identify and incorporate valid endpoints of TH action that would be equivalent to (e.g.,) seminal vesicle weight for androgens, represents a significant weakness in the EDSP that will create considerable debate about the interpretation of data derived from these assays.

2.7 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

Kevin Gaido: This report included multiple studies for reproducibility, sensitivity, and interlaboratory comparisons. The reproducibility and sensitivity studies included a number of different compounds known to impact androgen, thyroid or HPG function. Four chemicals were selected for the interlaboratory study. Dibutyl phthalate and vinclozolin were selected as antiandrogens, DE-71, a polybrominated diphenyl ether mixture was chosen to test thyroid-related endpoints. 2-CNB was chosen as a toxic compound that did not affect endocrine function. Unfortunately, this compound demonstrated endocrine disrupting activity and did not serve as a good negative control. The analytical and statistical methods appear appropriate.

Tom Zoeller: The test substances were appropriately chosen as chemicals that produced a relatively weak effect. This was a good test of the sensitivity of the assay and resulted in somewhat ambiguous results with Phenobarbital. The analytical methods were variable. Certainly body and organ weights are reasonably analyzed. However, histopathological analysis should be objectified with computer-assisted morphometry. Morphometry has been shown to be more reliable than visual estimation (1). In addition, the RIAs are problematic as has been discussed above. The description of the statistical methods is clear and appropriate.

2.7.1 Test Substances

George Daston: Fourteen chemicals were evaluated during the course of assay development and validation. The chemicals represent a range of modes of action, including androgens, anti-androgens, estrogens, thyroid-active compounds, and agents that affect the hypothalamic-pituitary-gonadal axis. In some instances chemicals with the same mode of action but different potencies were evaluated. The test substances were appropriate to evaluate the range of mode of actions which the assay is capable of detecting. Given the number of potential modes of action it would be very useful to have a much larger set of chemicals evaluated. However, I am aware of the number of other tasks that EPA needed to perform to validate other assays in the battery, and feel that the chemicals in the validation set for the pubertal male assay are representative of a wide range of modes of action and provided a rigorous test of assay performance. Only one presumably negative substance, chloronitrobenzene, was evaluated, and this compound had effects on assay endpoints. It will be important to evaluate assay performance with additional negative compounds.

Richard Dickerson: a wide ranging selection of test compounds was selected that included antiandrogens that ranged from weak to strong, compounds that had varied mechanisms of action on the androgen-pituitary-hypothalamus axis and the thyroid synthesis and metabolic pathways.

Richard Sharpe: The 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. This data has been obtained from several different but experienced laboratories. The main criticism is that no truly negative compound was evaluated; 2-CNB was chosen for this purpose, although I am frankly amazed that anyone would choose such a compound as a ’negative’ when closely related chemical compounds have been shown to be profound testicular toxicants in adult rats. Aside from this, the range of compounds chosen was wise and included a notably wide range of different hormonal activities. This choice undoubtedly proved that the assay is readily able to detect compounds that affect the ‘androgenic’ hormonal systems at various levels, and data for the

thyroid axis (in which I am inexpert) also appear reasonably convincing, with the notable exception of phenobarbital effects.

2.7.2 Analytical Methods

George Daston: The analytical methods were appropriate. Most of the endpoints are of organ weight and histology, which were straightforward, if somewhat variable for some tissues (e.g., ventral prostate). I believe that the variability will decrease as labs become more adept in the dissections and standardized in their procedures. The hormone measurements were conducted according to accepted methodology.

Richard Dickerson: Highly appropriate foe endpoints utilized in the assay.

Richard Sharpe: Essentially two analytical methods are used as part of the test, hormone assays and selective evaluation of organ histopathology (testes, epididymides, thyroid, kidney). For the most part, the hormone measurements do not constitute an important component of the pubertal rat assay, but if these are to be retained as part of the overall assay then standardization of the assay kits used is essential to provide uniformity as well as minimizing inter-laboratory variation. However, such variation is commonplace and likely to be considerable when, and if, the pubertal rat assay is put into widespread use by laboratories that have little experience with running hormone assays. In the validation studies, organ histopathology proved to be a rather insensitive endpoint as effects were only detected for compounds fully expected to have major target organ effects, namely DBP and 2-CNB on the testis and propylthiouracil, perchlorate and DE-71 on the thyroid. As each of these compounds had corresponding effects on organ weights and/or on relevant hormone levels, a case could be made for dispensing with organ histopathology in this Tier 1 assay, especially as it requires considerable histopathological expertise. Even though I am a great fan of histopathology, I am not convinced that it adds greatly to the assay, bearing in mind its primary objectives. I see no need for kidney histopathology.

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

George Daston: The analysis of interlab performance included a calculation of each lab's CV for the endpoints measured, which was an appropriate way of identifying robust endpoints and potential areas for improvement in assay performance. The graphical summaries of the CVs across laboratories in the interlab reproducibility study were useful in understanding the range of variability in the study. These analyses appear to have been appropriately done.

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.

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.

2.8 Provide 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: I believe that the results are promising. It is possible to run the protocol and obtain reasonably comparable results across laboratories and over time. Different labs were able to detect signals of endocrine activity, and it is unlikely that there would have been many, if any mistakes in the false negative direction had the assay been testing unknowns. There were differences among labs in the dose levels at which some effects were detected, which may have an influence on assay performance if dose selection is not perfect. The extent of variability for many of the endpoints is troubling: all labs were out of compliance with pre-set performance criteria for 4 of 17 endpoints for one lab, 5 of 17 for two, and 6 of 17 for one. In other words, roughly one-fourth to one-third of the endpoints were more variable than was believed to be

acceptable, a result that could compromise the resolving ability of the assay (as well as its reproducibility). These are issues that will need to be addressed in order for the assay to be used routinely to evaluate unknowns.

Richard Dickerson: Based upon the interlaboratory validation and transferability studies, the assay consistently gives results that are both reproducible and repeatable both within given laboratory and between laboratories with minor exceptions due to exceeding the CV for a number of endpoints. The only consistent endpoint that was exceeded the CV was determination of ventral prostate weight. These exceptions were mainly due to dissection technique differences. However, failure to keep the CV within the stated range did not prevent determination of an effect.

Kevin Gaido: There were inconsistencies in hormonal measurements between laboratories. This is likely due to biological variability but may also have to do with technique. Despite the inconsistencies the overall trend was consistent across laboratories and the redundancy of endpoints reduces concern regarding any one specific measurement. Thus, while there is some variability associated with specific endpoints in this assay, the inclusion of multiple endpoints increases its reliability.

Richard Sharpe: From the validation and other studies plus published studies, the reproducibility of the assay is impressive. From the inter-laboratory study involving two doses of each of four compounds (dibutyl phthalate, vinclozolin, 2-CNB and DE-61), the inter-laboratory reproducibility extended in almost all instances to both doses of each of these compounds. This was all the more impressive when considering that in many instances in the same comparison, most or all of the laboratories were unable to meet the CV performance criteria for these endpoints (discussed earlier). This imparts considerable confidence that the assay is inherently robust and reproducible and will be transferable between laboratories with relative ease; the comparatively simple format of the assay components reinforces this conclusion. The only outstanding issue is that of the false positive rate in the assay, but this should be resolvable in time with its wider application to chemicals with unknown activity.

Tom Zoeller: An important question is whether the variability in endpoints can be reasonably reduced – both within lab and between labs – by standardizing different elements of the test. A major variable will be that of the feed. We know that variation in the amount of estrogenic compounds in feed is high, regardless of the supplier's certification. This variation alone can interact with test compounds to provide variability from lab to lab, or at different times within the same lab. Variability in hormone levels will be affected by this, but also by the standardization methods as described above.

2.9 Additional Comments and Materials Submitted

Richard Sharpe: **Minor comment/correction**: On page 16, line 19 there is a statement to the effect that flutamide has been shown to 'feminize male external genitalia'. This is inaccurate. It is not strictly possible to 'feminize male external genitalia', which implies that there has been reversal of earlier masculinization (which has never been shown). Flutamide, and other anti-androgens, interfere with masculinization such that the genitalia are under-masculinized (ie. they do not shift from the 'set-up' female pattern).

Tom Zoeller:

1. **Hooth MJ, Deangelo AB, George MH, Gaillard ET, Travlos GS, Boorman GA, Wolf DC** 2001 Subchronic sodium chlorate exposure in drinking water results in a concentration-dependent increase in rat thyroid follicular cell hyperplasia. Toxicol Pathol 29:250-259

3.0 PEER REVIEW COMMENTS ORGANIZED BY REVIEWER

Peer review comments received for the male pubertal 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 Pubertal Development and Thyroid Function in Juvenile Male Rats as a Potential Screen in the Endocrine Disrupter Screening Program Tier-1 Battery

George Daston

October 8, 2007

1. 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. 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. 5, line 7). 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 line 8, p. 5. 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 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. The report does a good job of describing the selection of test chemicals for validation as being based on their mode of action. It should also be made explicit that successful validation would include results from the assay that correctly identify those modes of action.

The pubertal male assay is a complicated assay with a large number of endpoints. The purpose of the assay is also complex, with the intent to detect a large number of modes of endocrine action. The basic concept for the assay is that agents with a specific mode of action would affect only a subset of the endpoints measured, and that by analyzing which endpoints were affected it would be possible to discern mode of action. It would be very helpful to the reader to have a table early on in the report that describes which endpoints are expected to be affected for each mode of action being evaluated. This information is contained in the descriptions of the effects of the test agents used in the various studies, but it would be much easier to interpret the results if a summary table linking endpoints with mechanisms were provided.

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

The report describes in the context of the validation framework a large number of studies that were conducted for a variety of purposes. These include transferability of the assay, reliability, and validity of the assay for detecting several modes of action. For the most part, I found the

presentation to be clear and the interpretation consistent with the results of the studies. For the purpose of clarity in this review, I will address interpretation issues for each of the study types separately, and then the overall interpretation of the studies in aggregate.

Transferability study: This study was the first conducted outside of an investigative research lab and confirmed that the protocol yielded similar results in a separate laboratory. The study was extensive and evaluated six chemicals representing different modes of action, in two strains of rats. The results are thoroughly presented, and I agree with the overall interpretation that the protocol is transferable. That said, the claim of transferability would be better supported if data from similar studies (either from EPA's labs or from the literature citations described for each chemical) were presented along with the contract lab results. Transferability involves getting the same response (at least qualitatively) in two different labs. The text of the document indicates that this was the case, but it would help the reader to understand this if the original data were summarized in tabular form for easier comparison with the Therimmune data.

Page 12, lines 18-19, indicates that the research at Therimmune evaluated the transferability of the protocol "as it existed in 1999". It would be useful to describe briefly what the differences were so that the reader can be convinced that the conclusion of transferability is relevant to the current protocol. I assume that the difference was limited to a few endpoints that were not measured at the time (thyroid weight, testosterone) but that was not clear from reading the report.

The transferability study is also noteworthy in that it adequately addresses the question of strain sensitivity and supports the decision to use Sprague Dawley rats in subsequent studies.

Assay Sensitivity: This section describes studies to evaluate many more chemicals representing a range of modes of action, to evaluate multiple dose levels, and to assess the extent to which changes in body weight gain affect the other endpoints measured in the study. Again, the results are thoroughly described. I agree with most of the interpretation, but there are a few conclusions that are troublesome and that I believe should be looked at again.

The conclusion about phenobarbital (p. 32, lines 27-29) is that the protocol was sensitive to the thyroid-related effects of Phenobarbital despite the fact that there were no statistically significant

changes in most of the thyroid-related responses. The report blames the lack of significance on the fact that the MTD wasn't reached and that therefore the experiment was not an adequate test of the resolving power of the assay. I can accept that rationale, but if true, then nothing can be said about the ability of the assay to detect Phenobarbital as an indirect thyroid toxicant. The preceding text on p. 32 is clear about that point, but it should be reflected in the concluding paragraph.

The second result that deserves some additional discussion is the number of compounds that affect thyroid hormone (T4) concentrations. It appears that T4 is not a particularly specific measure of thyroid toxicity. This lack of specificity has been noted in other screening protocols (e.g., O'Connor et al., 2002) and should be discussed in this report in some depth, either in this section or in the overall conclusions. It is possible that T4 measurements can only be interpreted in the context of some other effect to be meaningful. The more that this is discussed before this or other assays come on line, the better the interpretation of results will be for compounds with uncharacterized modes of action.

The multi-dose study is thoroughly described. There is only one point on which I believe the interpretation should be expanded: p. 36, lines 4-7, it is explained that an apparent result of flutamide on adrenal weight was probably due to inordinately low values in two control animals and was not an effect of the compound. I agree with this conclusion, but it raises the question of whether this apparent variability problem is important enough to correct. For example, in another part of the report, it was determined that high variability in the weights of fluid-filled organs indicated that more detail and training was needed across labs in proper dissection. In this instance, could it also be a problem with procedure, or is the variability due to the animals themselves, in which case it may be important to either increase Ns or to amass a larger historical control data set against which new results can be compared.

The section describing the study on the effects of body weight gain is satisfactorily described. There is one semantic correction that I would like to see made in this section, which is to change all references to "body weight loss" to "decreased body weight gain". The latter is actually what was measured. The animals on restricted feed were not losing weight; they were gaining weight

at a slower pace than controls. This distinction is important because a misunderstanding of this point could lead to inappropriate dose selection, excessive toxicity, and an uninterpretable study.

The final paragraph of this section (p. 47-48) needs to be clarified or expanded on. The first sentence of the paragraph states that a 10% reduction in body weight is a reasonable basis for setting an upper limit in the assay. However, the next sentence suggests that this level of effect is too severe for thyroid effects and the final sentence indicates that a 9-10% decrease may necessitate conducting additional studies and/or a weight-of-evidence approach to interpret the results for the thyroid endpoints. This seems to indicate that the male pubertal assay cannot be optimized for thyroid endpoints and reproductive endpoints, and that either a compromise would need to be made in selecting an upper dose level, or an additional assay would have to be run. This is an important point in determining whether the assay is a suitable alternative, and should be discussed at greater length.

Performance Criteria: The section on performance criteria is adequately presented, but Table 17 needs some attention. The column heading "1.5 CV" must be incorrect (since the values are less than those for CV in the column immediately to the left). I believe it should be "1.5 SD". Also, the last column would be more understandable if it were titled something like "Maximum acceptable CV".

Interlaboratory study to examine reproducibility: This section adequately describes the study and its results. There are only a few points that I found troubling. First, on p. 52, lines 14-18, it is stated that EPA does not require the assay to consistently display a pattern of endpoint responses diagnostic for a particular mode of action, but only that thyroid responses not be used to claim consistency with sex steroid associated responses and vice versa. Given that the stated purpose of the assay is to detect a variety of hormone-related modes of action, and that, as the most apical assay in the screening battery, it will have the greatest influence on weight of evidence determination of the battery, we need to expect more of the assay.

The final paragraph on p. 72 indicates that, while the assay results are generally reproducible, the CV criteria that were established a priori were not always met. I don't see this as a fatal flaw with the assay, but it will be important in the early going to constantly re-evaluate the magnitude

and sources of variability and to find ways to minimize the latter and better set acceptable criteria for the former.

3. 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 IV (p. 9). I believe that this assay, as well as the adult male and pubertal female assays being evaluated, has the potential to provide the most reliable and comprehensive information for the weight-of-evidence determination. 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 pubertal male also has special relevance in that puberty represents a major developmental stage in the maturation of the reproductive system, and may be particularly susceptible to exogenous agents that interfere with the hormonal control of sexual maturation. Because young, rapidly growing animals are used, the system is expected to be sensitive to agents that affect thyroid function. Stoker et al. (2000) provides an excellent review of the literature on the effects of exogenous agents on puberty and thyroid function in the pubertal rodent. This review provides a strong theoretical underpinning of the relevance of the assay for the stated purpose of screening for potential effects on androgen and thyroid-dependent systems, and possibly other modes of action.

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. Clarity and conciseness of the protocol in describing the methodology of the assay such that the laboratory can:

a. comprehend the objective

This section is succinct and to the point.

b. conduct the assay

The protocol contains enough information for a competent laboratory to conduct the assay in a consistent way. There are a few aspects of the protocol that may be too restrictive, such as the admonition to keep temperature and relative humidity within ranges that may not be achievable in all facilities, and are different from those established by AAALAC.

c. observe and measure prescribed endpoints

The information in the protocol and attachments were clear and helpful in providing guidance on evaluating the endpoints. I found it very useful that the attachments to the protocol included information useful to conducting the assay, such as reference images for thyroid histology.

d. compile and prepare data for statistical analysis

Sufficient guidance was provided on how the data should be displayed and analyzed. The statistical procedures were not overly restrictive but provided enough guidance to facilitate comparison of study results in different labs.

e. report results

The protocol provided a great deal of information on how to report and interpret results. The length of the data interpretation section is unusual in my experience, but given the novelty of the protocol and the complexity of interpretation, I believe it to be warranted.

5. Strengths and/or limitations of the assay in the context of a potential battery of assays to determine interaction with the endocrine system.

The section on pp. 90-93 of the report does a good job of summarizing the strengths and limitations of the assay. I consider the major strengths of the assay to be:

a. The assay is performed in an intact system, capable of responding to interactions among multiple endocrine axes.

b. The assay uses developing animals, which may be more susceptible to endocrine perturbations, thereby potentially increasing the sensitivity of the assay.
c. The assay measures endpoints indicative of a number of modes of action of interest to EPA for endocrine screening. It should be possible to obtain a lot of information out of a relatively few animals.
d. The apical nature of the assay will allow it to provide a lot of context for weight-of-evidence interpretation of some of the simpler assays in the proposed tier 1 battery. It also allows for multiple endpoints for each mode of action, improving the interpretability of the assay.

The limitations of the assay are

a. The assay is lengthy. Dosing covers a 30 day period. Furthermore, the study requires a pilot study for dose-setting. The length of the assay is not conducive for rapid screening.
b. The assay appears to be inordinately influenced by changes in body weight gain, such that a significant change in body weight gain is needed for a valid (presumably, sensitive) assay, but changes above 9-10% make the assay difficult to interpret. It may not be possible to achieve this much precision in dose setting, especially in a screening context with limited numbers of animals.
c. There is not enough information to assess the specificity of the assay. It may be too non-specific to provide enough information to correctly classify negatives. More research needs to be done to address this concern.
d. There is not enough information yet to make definitive conclusions about the range of modes of action that are detectable by the assay. The evidence is good that the assay can detect anti-androgens and thyroid-active agents. There is also reasonable evidence that the assay can detect androgens, agents that affect the hypothalamic-pituitary gonadal axis, and steroidogenesis inhibitors. The evidence on estrogens and aromatase inhibitors is mixed. These may not be limitations of the assay, but are limitations on how we can interpret the utility of the assay at present. Clarity on which modes of action it covers will be important in determining the assays that should be conducted with it as part of a battery.

e. The apparent non-specificity of T4 is concerning, and may suggest that this measure cannot be interpreted out of context with other thyroid measures.

6. Impacts of the choice of:

a. test substances

Fourteen chemicals were evaluated during the course of assay development and validation. The chemicals represent a range of modes of action, including androgens, anti-androgens, estrogens, thyroid-active compounds, and agents that affect the hypothalamic-pituitary-gonadal axis. In some instances chemicals with the same mode of action but different potencies were evaluated. The test substances were appropriate to evaluate the range of mode of actions which the assay is capable of detecting. Given the number of potential modes of action it would be very useful to have a much larger set of chemicals evaluated. However, I am aware of the number of other tasks that EPA needed to perform to validate other assays in the battery, and feel that the chemicals in the validation set for the pubertal male assay are representative of a wide range of modes of action and provided a rigorous test of assay performance. Only one presumably negative substance, chloronitrobenzene, was evaluated, and this compound had effects on assay endpoints. It will be important to evaluate assay performance with additional negative compounds.

b. analytical methods

The analytical methods were appropriate. Most of the endpoints are of organ weight and histology, which were straightforward, if somewhat variable for some tissues (e.g., ventral prostate). I believe that the variability will decrease as labs become more adept in the dissections and standardized in their procedures. The hormone measurements were conducted according to accepted methodology.

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

The analysis of interlab performance included a calculation of each lab's CV for the endpoints measured, which was an appropriate way of identifying robust endpoints and potential areas for improvement in assay performance. The graphical summaries of the CVs across laboratories in the interlab reproducibility study were useful in understanding the range of variability in the study. These analyses appear to have been appropriately done.

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

I believe that the results are promising. It is possible to run the protocol and obtain reasonably comparable results across laboratories and over time. Different labs were able to detect signals of endocrine activity, and it is unlikely that there would have been many, if any mistakes in the false negative direction had the assay been testing unknowns. There were differences among labs in the dose levels at which some effects were detected, which may have an influence on assay performance if dose selection is not perfect. The extent of variability for many of the endpoints is troubling: all labs were out of compliance with pre-set performance criteria for 4 of 17 endpoints for one lab, 5 of 17 for two, and 6 of 17 for one. In other words, roughly one-fourth to one-third of the endpoints were more variable than was believed to be acceptable, a result that could compromise the resolving ability of the assay (as well as its reproducibility). These are issues that will need to be addressed in order for the assay to be used routinely to evaluate unknowns.

3.2 Richard Dickerson Review Comments

CHARGE QUESTIONS

Each peer reviewer is asked to review the Integrated Summary Report and protocol (appendix 1), and comment on the results of the validation process of the male pubertal assay, especially the results of the interlaboratory validation exercise. Review and comment shall be directed to each of the following:

1. Clarity of the stated purpose of the assay.

   The first paragraph of the stated purpose of the assay is clear enough but could be improved by eliminating or replacing some of the phrases. For example, the phrase "information that will be useful in assessing the potential of a chemical substance or mixture to interact with the endocrine system" is much too long and passive. Consider

replacing it with “information useful in determining the potential of chemicals or mixtures to interact with the endocrine system.

In the second paragraph, it is stated that the test compounds will be dissolved in corn oil whereas test substances are dissolved in a methyl cellulose vehicle for the 15-day adult male assay. Why the difference in choice of vehicles? Corn oil contains varying amounts of phytoestrogens that may mask some changes in the endocrine system. It seems like a more inert vehicle such as triolean would be more appropriate. In addition, pups should be culled to 8 or 10 per litter rather than the range of 8-10 and housed 2 or 3 per cage instead of 2-3 to further reduce variability. There is no strain of rat listed in this brief description of the assay; shouldn’t it be stated here?

2. 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. Several areas of concern were discussed including strain variability, variability in determining the day of PPS, and measurement of weight of fluid-filled organs. However, reasonable solutions were proposed for each of these areas of concern.

For the multiple chemical studies, the effects of phenobarbital exposure on the thyroid axis were noted as non-significant but the trends were in the correct direction. One reason suggested for the lack of significance was failure to reach the MTD. Failure to reach the MTD may be common when there is a lack of data concerning the toxicity of test compounds. Perhaps a quantitative method of determining significance when a number of endpoints almost reach the stated level of significance could be used or the standard could be multiple endpoints reaching the 0.10 level of significance as a way of determining weight of evidence.

3. 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. These endpoints are sensitive to exposure of known androgen and thyroid agonists and antagonists. The endpoints used for the HPT axis are also the most appropriate for the length of the assay. Therefore, the assay measures physiological endpoints appropriate for detecting alterations in the status of the male reproductive and thyroid organ systems.

In terms of toxicologic relevance, the endpoints selected for the Pubertal Male Rat Assay are appropriate for several reasons. First, they reflect biologically relevant endpoints as discussed above. Second, validation 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. Third, exposure to a dopamine antagonist, pimozide, showed that the assay was sensitive to compounds that inhibit prolactin release. Fourth, exposure of test animals to propylthiouracil, a thyroid hormone synthesis inhibitor, and to phenobarbital, which increases metabolism of thyroid hormones, demonstrated the assay can detect compounds that alter the production or clearance of thyroid hormones. 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. 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 attached as Appendix 1.

    a. comprehend the objective- Section I. Purpose and Applicability is clearly worded and states the purpose of the assay precisely in such a way that individuals with scientific

training can easily comprehend it. It might require re-phrasing for a less technically audience or addition of a lay summary.

b. conduct the assay- The instructions on how to conduct the assay are complete and clear for the most part. Prohibition of the use of tap water for the animals is very good to eliminate perchlorates and other contaminants. Specifying rat chow that is low in phytoestrogens is good but consider adding as an appendix a list of acceptable rodent chows. Replace the words kill, killed or killing with euthanized or euthanasia. The humidity conditions are perhaps too stringent and may prevent the assay from being performed in regions of the country that are too dry or too humid. For example, facilities in West Texas, New Mexico and Arizona frequently can not maintain 30% humidity. Section III states that the litters will be standardized or culled to 8-10 pups per litter. For a particular study, all litters should be standardized to the same size. Why is it not acceptable to cross-foster to raise the litter size to the minimum? It can prevent waste of animals if used properly. Section VI discusses the properties of corn oil to be used but does so in qualitative terms. Consider changing it to corn oil from an approved source, from a freshly opened container, and free of sediment. However, the use of corn oil is problematic in that it can contain phytoestrogens. Why not use an synthetic oil such as triolean? Section IX needs to be revised to meet AVMA Panel on Euthanasia standards. See the note below. The remainder is fine.

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.

Notes:

1. Decapitation without prior use of either CO2 or inhalational anesthetic is not an AVMA Panel on Euthanasia (2000) approved method of euthanasia. It is listed as a conditionally acceptable method by the Panel but requires justification for use such as peer-reviewed journal articles that conclude prior exposure to CO2 or inhalational anesthetic alters the endpoints that are the focus of the study. The statement is made in the protocol that decapitation is considered more humane than CO2 asphyxiation. This is

in direct opposition to the AVMA Panel on Euthanasia which is reference used by IACUCs, AAALAC, the NIH Guide, and the USDA.

2. Heat-treated aspen shavings are preferred to heat-treated pine shavings since they contain lower levels of volatile compounds and appear to be less allergenic to workers.

5. Strengths and/or limitations of the assay in context of a potential battery of assays to determine interaction with the endocrine system.

Strengths of the assay include ease of conducting the assay and measuring the endpoints, the moderate duration of exposure, biological and toxicological relevance and substantial literature base. In addition, this assay uses intact animals whereas other assays such as the Hershberger assay require castrated animals. Furthermore, many of the techniques used are common to other assays such as the 15-day adult rat assay, multigenerational testing protocols, and Hershberger assays that have been used for some time. The use of time of PPS as an endpoint is an advantage due to its increased sensitivity to androgen agonists and antagonists compared to weight of androgen-responsive organs and tissues. Coupling this assay with the pubertal female rat assay provides a robust test capable of detecting xenoestrogens, antiestrogens, androgen agonists and antagonists as well as inhibitors of thyroid synthesis, inducers of thyroid hormone metabolism and compounds that alter release of pituitary hormones.

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. This assay does not differentiate between the various mechanisms of action by which a compound can affect androgen status or whether change in thyroxine is due to inhibition of synthesis or induction of metabolism. This assay also uses timed pregnant rats which are not only live animals but fairly expensive. Although the exposure period is short in comparison to some other assays, it does require 30 days. Last, this assay is more recent than the Hershberger assay or the rat uterotrophic assays which translates into a smaller data base and less harmonization.

6. Impacts of the choice of:

   a. test substances- a wide ranging selection of test compounds was selected that included antiandrogens that ranged from weak to strong, compounds that had varied mechanisms of action on the androgen-pituitary-hypothalamus axis and the thyroid synthesis and metabolic pathways.
   b. analytical methods- highly appropriate foe endpoints utilized in the 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. Repeatability and reproducibility of the results obtained with the assay, considering the variability inherent in the biological and chemical test methods.

Based upon the interlaboratory validation and transferability studies, the assay consistently gives results that are both reproducible and repeatable both within given laboratory and between laboratories with minor exceptions due to exceeding the CV for a number of endpoints. The only consistent endpoint that was exceeded the CV was determination of ventral prostate weight. These exceptions were mainly due to dissection technique differences. However, failure to keep the CV within the stated range did not prevent determination of an effect.

3.3 Kevin Gaido Review Comments

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

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

The stated purpose of the male pubertal assay, is to provide an alternative to the female pubertal assay as an in vivo mammalian system useful for detecting chemicals that interfere with androgen or thyroid function, or alter hypothalamic function, gonadotropin or prolactin secretion.

The assay will ultimately become part of a comprehensive battery of tests for endocrine active chemicals.

2. *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 available data. A comprehensive review of the literature is presented and considered. A detailed comparison of the results from each laboratory together with historical data is provided. Results from the interlaboratory validation demonstrate that the protocol is transferable and reproducible and capable of detecting chemicals that act through a variety of endocrine related mechanisms to impact male pubertal development.

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

The assay was designed to detect chemicals that interfere with androgen or thyroid function or with the HPG axis based on the understanding of the biological relevance of these functions for normal pubertal development. Serum hormones and reproductive organ weights significantly increase in male rats during puberty and as a result, chemicals that disrupt endocrine function can have a dramatic impact on male pubertal developmental measurements such as organ weights and preputial separation. This assay is highly relevant for toxicological screening for endocrine active chemicals.

4. *Clarity and conciseness of the protocol in describing the methodology of the assay.*

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. Methods for housing and treatment of the animals are presented in sufficient detail. Each endpoint is clearly described and methods for statistical analysis as well as how to handle outliers are presented. Finally details and examples for data interpretations, presentation, and developing a final report are given. The reproducibility and transferability of the assay is clearly demonstrated by the reproducibility of overall results across laboratories. While there was some variability with some endpoints between the laboratories the overall weight of evidence and conclusions were consistent.

5. *Strengths and/or limitations of the assay in the context of a potential battery of assays to determine interaction with the endocrine system.*

Strengths of this assay include the ability to screen for multiple modes of action in a sensitive *in vivo* mammalian assay. The assay has extensive historical data from multiple laboratories and the biology behind the various endpoints is well understood. The assay focuses on a time period when reproductive organ development is very sensitive to endocrine disruption. 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. The assay has multiple and redundant sensitive endpoints that can be used to help design more definitive Tier-2 testing. For example, results with 2-CNB provided sufficient information to suggest that its effect on the growth of androgen dependent tissues is either through altering steroidogenesis or targeting the secretion of pituitary hormones but not through interference with the androgen receptor.

A weakness of this assay may be its ability to screen for week estrogens. While the assay can detect more potent estrogens, there is insufficient data regarding weaker estrogens. In addition, the lack of a test of a negative control makes it difficult to determine the specificity of this assay for endocrine active compounds. The reliance on MTD is a weakness since additional prior studies must be performed to accurately identify MTD or in cases where the MTD is based on a review of the literature may lead to over or underestimating the MTD.

6. *Impacts of the choice of test substances, analytical methods, and statistical methods in terms of demonstrating the performance of the assay.*

This report included multiple studies for reproducibility, sensitivity, and interlaboratory comparisons. The reproducibility and sensitivity studies included a number of different compounds known to impact androgen, thyroid or HPG function. Four chemicals were selected for the interlaboratory study. Dibutyl phthalate and vinclozolin were selected as antiandrogens, DE-71, a polybrominated diphenyl ether mixture was chosen to test thyroid-related endpoints. 2-CNB was chosen as a toxic compound that did not affect endocrine function. Unfortunately, this compound demonstrated endocrine disrupting activity and did not serve as a good negative control. The analytical and statistical methods appear appropriate.

7. *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. This is likely due to biological variability but may also have to do with technique. Despite the inconsistencies the overall trend was consistent across laboratories and the redundancy of endpoints reduces concern regarding any one specific measurement. Thus, while there is some variability associated with specific endpoints in this assay, the inclusion of multiple endpoints 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 somewhat arbitrary as in some instances it is equivocal as to which of the questions posed they address. I have little expertise with regard to the thyroid axis, so have restricted my comments on this aspect, and evaluation by those with proven expertise on this axis should be heeded rather than my own comments in this regard.

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

   The background information and protocol description give a clear view of the objectives of the assay and the role of its component parts. This material should be easily comprehensible to anyone intending to use and apply this assay. It is a Tier-1 assay and its priority is to maximize the detection of endocrine active compounds (that target the male reproductive and/or thyroid axes) whilst minimizing false negatives. The use of multiple endpoints (mainly organ weights) is designed to ensure this. The use of an intact animal with normally functioning, homeostatic, hormone systems is, in my opinion, a strength as it represents the 'real world'. It includes metabolism etc and can potentially 'integrate' chemical effects that may directly affect one hormonal (or non-hormonal) system and then secondarily affect another hormonal system.

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

A considerable amount of data has been accrued on this assay and has involved several different (but experienced) laboratories and the testing of a large number of compounds with a wide variety of purported or known mechanisms of action (MOA). The evidence presented for review and in a few publications substantiate the view that this assay is fit for purpose. As with the intact adult rat assay, the pubertal rat assay uses multiple endpoints in order to maximize detection of 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). It is simpler than the adult male assay in terms of endpoints (less hormonal data), and arguably this makes the assay potentially more straightforward to operate and interpret, though it requires a considerably longer treatment period. The use of multiple endpoints may provide preliminary information on the potential MOA but, 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 several inter-laboratory comparisons, are rational and fit with current understanding of how the various endocrine systems operate within the body during pubertal development. There are some issues in relation to assay specificity and which endpoints are absolutely essential, and others which might be dispensable or assigned a ‘supporting role’ (see below), but these are rather minor issues, and they do not affect the overall conclusion that the assay appears robust and fit for purpose, but with some limitations.

Although results obtained in applying the assay to a variety of chemicals with known endocrine activity have been largely as expected, the specificity of the assay is to some extent unproven. In the inter-laboratory study, 2-chloro nitrobenzene (2-CNB) was included as a test of specificity as various indicators had suggested that it would not have ‘endocrine activity’ and would thus provide a test of assay specificity. However, 2-CNB significantly reduced testosterone levels, weights of all androgen-dependent organs and delayed preputial separation as well as causing histopathological changes to the testes. Based on these findings, 2-CNB would

clearly be classed as an endocrine disruptor and it must be presumed that this is the case (nitrobenzene and dinitrobenzene are well-established testicular toxicants in the adult rat, although I am not aware of evidence that they disrupt Leydig cell function). The alternative interpretation is that the pubertal assay is prone to non-specific effects and will therefore yield 'false positives' with a high frequency. Although it is preferable for a Tier-1 (screening) assay to suffer from this rather than from frequent false negatives, it will be important in future studies to more rigorously investigate the specificity of the pubertal assay. My expectation is that the assay will not be unduly prone to false positives, based on the relative lack of impact of the feed restriction studies on the assay endpoints, but this needs to be demonstrated categorically using appropriate test compounds.

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

The validation process, which involved testing of a considerable number of compounds with a range of known activities and mechanisms of action provided solid foundations for subsequent evaluation and interpretation of results for compounds of unknown hormone activity. It also identified unexpected effects or results, such as those for 2-CNB and phenobarbital, which are discussed elsewhere. It is likely that continued application of the assay to a wider range of chemicals will uncover other activity profiles that do not fit within our expected concepts, but this will inevitably lead to a better understanding of the utility, and limitations, of the assay. I anticipate a progressive ability to categorize chemicals into classes based on their activity profile in this assay, even if it is not possible to define a clear MOA. Because the test uses an intact animal that is advancing through one of the most endocrinologically dynamic phase of its life (puberty), it might be anticipated that compounds might affect target organs or hormone levels via pathways that are unrelated to endocrine disruption *per se*, for example effects on food intake/metabolism that lead secondarily to such changes. However, from results obtained so far, and including food restriction studies, this expectation has not surfaced in any major way, which arguably makes the assay more robust, simpler in its function and analysis and relatively easy to interpret.

4. **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 competent to judge (as a scientist running an academic research laboratory), the protocol provided is clearly laid out, is understandable and is sufficiently detailed to enable an appropriately experienced laboratory to run, complete, evaluate and report results using this assay. There are no deficits in the protocol that I have noticed. The fact that only 3 hormonal assays are included, namely for thyroxine (T_4), thyroid stimulating hormone (TSH) and testosterone, makes the pubertal assay relatively easy to run, as hormone assays are often a source of considerable inter-laboratory variation and, because they provide only a 'snap-shot' in time of hormone levels, they can be potentially misleading.

Arguably one of the more variable and more subjective aspects of the pubertal assay is the endpoint of preputial separation (PPS). This is clearly a useful 'endocrine' endpoint that summates androgen action over a period of time, as shown in all of the studies so far done to test and hone the pubertal assay. Nevertheless, in the inter-laboratory study none of the 4 (experienced) labs involved could meet the coefficient of variation (CV) criterion for bodyweight at PPS and only 2 could do so for age at PPS. Some refinement of the PPS 'definition' was adopted after issues relating to incomplete PPS (retention of 'threads' of connecting tissue), but by the very nature of the assessment it seems to me that PPS will always be prone to high between-laboratory and between-observer variation. Similar to PPS, the recorded weights of the ventral prostate, seminal vesicles + coagulating gland, epididymis and levator ani +bulbocavernosus muscle all provide a measure of androgen action over time – in essence they summate androgen action over the 30-day course of the assay. Accordingly, weights of these organs changed more or less in parallel to PPS in response to exposure to the various compounds tested in the different parts of the validation exercise and in other studies. My initial reaction to this (expected) observation was that PPS was maybe redundant, and could therefore be dispensed with, as it did not measure anything that the target organs already did not. However, it is also apparent that for most of these organ weights there is similarly high CV as discerned from the inter-laboratory study. Thus, none out of the 4 labs met the CV criterion for ventral prostate weight, only 1 met the CV for

seminal vesicle weight and only 2 did so for epididymal weight. None of this is unduly surprising, as anyone experienced in the dissection of these organs will know that not only is there high variability in actual weight (probably largely reflecting different levels of fluid content), but the dissection process can also be variable depending on how this is done. For this reason it is good practice to have the same person do all of the dissections for the same organ in order that variation within a laboratory is minimized.

Returning to the issue of whether PPS is worth retaining, I convinced myself that it was, based on two lines of reasoning. First, it is an 'in-life' measurement, and thus may provide the first indication of 'anti-androgenic' or 'androgenic' activity in a test compound which can then be confirmed by organ weight measurements. Second, as PPS and reproductive organ weights are all intrinsically highly variable measurements (for reasons outlined above), it is safer to have multiple endpoints that reflect the same underlying phenomenon/activity (ie. androgen action over time), as this will increase the chances of detecting a significant effect on any one of the endpoints; the fact that one is in-life and the others terminal reinforces this argument. Additional to this reasoning is that PPS is non-invasive and not time-consuming as the visual inspection can be made at the same time as dosing the animal.

5. Strengths and/or limitations of the assay

The four main strengths of the assay are: (1) that it has a strong foundation based on various studies that have been undertaken as part of the validation exercise using a variety of compounds with different activities; (2) the primary reliance on target organ weights, which are easily measured, as the most predictive endpoints; (3) the use of multiple endpoints to inform on the same 'hormonal activity', some of which extend beyond organ weights (histopathology, hormone levels) and one of which is a dynamic pubertal 'in-life' event measured over time (PPS); (4) the minimal use of hormone measurements, as these are not only more technically difficult to measure, but may be difficult to interpret at a time in development when there can be large fluctuations within, and especially between, animals.

The main weakness of the assay, which at this stage of evaluation is more theoretical than experience-based, is that exposure of rats to compounds during such a hugely dynamic phase of growth and reproductive development as puberty is likely to result occasionally in effects

on reproductive ‘endocrine systems’ that are not due to intrinsic ‘endocrine activity’ but to an indirect effect eg. on growth. Although the feed restriction studies did not produce evidence to suggest major impact of this on reproductive organ weights or PPS in the pubertal assay, I still consider this to be a likely event with some compounds; the effects of phenobarbital are worth discussing in this respect (see below). Nevertheless, such an outcome will mean at the worst that some compounds are identified as ‘false positives’ which is acceptable in any screening assay provided this is not unduly frequent; the issue of assay specificity, which remains to be resolved, is important in this regard and has been discussed above.

I was somewhat disconcerted to see that, in both the multi-chemical study and in the TherImmune2 study, phenobarbital was detected as a clear anti-androgen, based on delayed PPS and reduced reproductive organ weights; equally disconcerting was its lack of significant effect on the thyroid axis. My presumption was that phenobarbital had been included as a test compound primarily because of its potential thyroidal effects (clearly evident in the adult male rat assay) and because its central effects might also ‘spill over’ into effects on the reproductive axis. As there is published data to show that phenobarbital can suppress LH secretion and even inhibit (fetal testis) steroidogenesis, the effects in the pubertal assay are not entirely unexpected but I was surprised at their robustness. It is possible that the pubertal rat is especially vulnerable to central effects of phenobarbital as normal progression of puberty is dependent on a progressive increase in frequency and amplitude of LH pulses, which then drive increasing testosterone production; the alternative interpretation, namely that the pubertal rat is more susceptible to non-specific effects remains to be resolved, as discussed above.

6. Comments on the impacts of the choice of:

a. test substances

The 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. This data has been obtained from several different but experienced laboratories. The main criticism is that no truly negative compound was evaluated; 2-CNB was chosen for this purpose, although I am frankly amazed that anyone would choose such a compound as a ’negative’ when closely related chemical compounds

have been shown to be profound testicular toxicants in adult rats. Aside from this, the range of compounds chosen was wise and included a notably wide range of different hormonal activities. This choice undoubtedly proved that the assay is readily able to detect compounds that affect the 'androgenic' hormonal systems at various levels, and data for the thyroid axis (in which I am inexpert) also appear reasonably convincing, with the notable exception of phenobarbital effects.

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, kidney). For the most part, the hormone measurements do not constitute an important component of the pubertal rat assay, but if these are to be retained as part of the overall assay then standardization of the assay kits used is essential to provide uniformity as well as minimizing inter-laboratory variation. However, such variation is commonplace and likely to be considerable when, and if, the pubertal rat assay is put into widespread use by laboratories that have little experience with running hormone assays. In the validation studies, organ histopathology proved to be a rather insensitive endpoint as effects were only detected for compounds fully expected to have major target organ effects, namely DBP and 2-CNB on the testis and propylthiouracil, perchlorate and DE-71 on the thyroid. As each of these compounds had corresponding effects on organ weights and/or on relevant hormone levels, a case could be made for dispensing with organ histopathology in this Tier 1 assay, especially as it requires considerable histopathological expertise. Even though I am a great fan of histopathology, I am not convinced that it adds greatly to the assay, bearing in mind its primary objectives. I see no need for kidney histopathology.

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

   From the validation and other studies plus published studies, the reproducibility of the assay is impressive. From the inter-laboratory study involving two doses of each of four compounds (dibutyl phthalate, vinclozolin, 2-CNB and DE-61), the inter-laboratory reproducibility extended in almost all instances to both doses of each of these compounds. This was all the more impressive when considering that in many instances in the same comparison, most or all of the laboratories were unable to meet the CV performance criteria for these endpoints (discussed earlier). This imparts considerable confidence that the assay is inherently robust and reproducible and will be transferable between laboratories with relative ease; the comparatively simple format of the assay components reinforces this conclusion. The only outstanding issue is that of the false positive rate in the assay, but this should be resolvable in time with its wider application to chemicals with unknown activity.

Minor comment/correction: On page 16, line 19 there is a statement to the effect that flutamide has been shown to ‘feminize male external genitalia’. This is inaccurate. It is not strictly possible to ‘feminize male external genitalia’, which implies that there has been reversal of earlier masculinization (which has never been shown). Flutamide, and other anti-androgens, interfere with masculinization such that the genitalia are under-masculinized (ie. they do not shift from the ‘set-up’ female pattern).

3.5

Thomas Zoeller Review Comments

REVIEW:

VALIDATION OF A TEST METHOD FOR ASSESSMENT OF PUBERTAL DEVELOPMENT AND THYROID FUNCTION IN JUVENILE MALE RATS 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)]*. One of the test systems recommended by the EDSTAC was the male pubertal rat assay. The purpose of the pubertal assay is to provide information obtained from an *in vivo* mammalian system that will be useful in assessing the potential of a chemical substance or mixture to interact with the male endocrine system. This assay is capable of detecting chemicals with antithyroid, androgenic, or antiandrogenic [androgen receptor (AR) or steroid-enzyme-mediated] activity or agents which alter pubertal development via changes in gonadotropins, prolactin, or hypothalamic function. 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 a well-written protocol and a well-written and well-managed validation study. Clearly, the EPA thought about performance criteria and about the logic required to interpret the findings. The performance criteria should be sharpened, both or the way the endpoints at necropsy are evaluated as the RIA performance criteria. There are additional comments, aside from the responses to the charge questions, that will be reviewed first.

General Comments:

Page 9, line 3, "…but these were later removed from the protocol as being relatively uninformative due to wide variation in levels."

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.

Page 24, line 7, "Although the primary cellular mechanism for this compound's effects on endocrine function are not characterized, it is well established that atrazine disrupts the hypothalamic (central nervous system, CNS) control of pituitary function by suppressing the gonadotropin releasing hormone (GnRH) stimulation". This statement does not appear to be referenced. Certainly, the work by Cooper et al., (1999 and 2000), and the work of McMullin et al. (2003) cannot be used to support this statement.

Page 25, lines 13-24, "…The conclusion from this study was that the pubertal male assay clearly identified atrazine as interacting with the endocrine system at both dose levels, thus showing that the assay is sensitive to chemicals that affect the HPG axis….". EPA's conclusion is dangerous! Rather, this study shows that when the EPA has previous information indicating that a chemical interacts with the endocrine system, they can selectively interpret the data in a way that is consistent with what is known. It is important to recognize that this entire paragraph amounts to

arm waving. It would be interesting to see what would happen if the EPA were to test chemicals in a blinded study in which neither the contract lab, nor the EPA interpreters were aware of the test chemical.

Page 26, lines 9-14. This conclusion ignores the observation made directly above it that serum T_4 levels were reduced, although serum TSH was not altered, nor were there treatment-related histological changes in the thyroid gland. No basis is given for ignoring T_4 levels.

Page 28, lines 8-12. This is a rational statement, but the logic is not spelled out.

Page 28, lines 14-18. This paragraph illustrates a weakness in the EPA's logic. First, low thyroid hormone can cause a decrease in weight gain; thus, animals treated with high levels of PTU could have a lower body weight precisely because serum T_4 levels are lower. In addition, serum T_4 levels are sensitive to caloric restriction; therefore, if animals treated with a high dose of compound such that caloric intake is restricted, serum T_4 levels could be lower due to this mechanism.

Page 31, lines 20-22. "Thyroid weight was increased at both doses, though the increases were not statistically significant (27.3 mg in controls, 31.9 and 32.5 mg at the low and high dose levels, respectively). As a fundamental rule, if something is not statistically significantly different, it is not different. Also as a fundamental rule, the biological significance of a difference may be arguable, but if an endpoint is statistically significantly different, it should be reported and interpreted as such; and if not….

Page 32, Lines 27-31. "The conclusion of this part of the study was that the male pubertal protocol appears to be sensitive to the thyroid-related and gonadal effects of Phenobarbital even though the thyroid-related responses were not significant at the $p<0.05$ level." This exact same profile was observed with Linuron and with Flutamide, yet the EPA concluded that the thyroid endpoints were not important. ***These studies are showing only that EPA can identify a well-known endocrine disruptor, not that they can identify an endocrine disruptor for which there are no previous data***.

Page 34, lines 10-11. "Due to an oversight, serum hormone levels (T_4, TSH, testosterone) were not obtained in this study." This demonstrates that "GLP" means only that record keeping is

precise, not that the study was performed according to plan, or that the techniques used to perform the study were appropriate or adequate.

Page 39, lines 5-7. "…a chlorotriazine herbicide which had recently…." This statement lacks foundation. The paper by Stoker et al. 2000a does not show this. My search of MEDLINE did not reveal studies that would support this statement. My query of investigators in this field also did not reveal studies that would support this statement.

Page 40, line 10-12. "At this point, no environmental chemicals have been found to bind to the thyroid receptor. (See Stoker et al. (2000b) for review of toxicant effects on thyroid function.)" **This statement is incorrect. Several compounds have now been shown to bind to the TR, some with IC50's in the nM range. The authors are not even using the EPA-managed thyroid DRP.**

Page 50, Table 17. It is not clear how these data were derived.

Page 51, Table 17. The performance standards for the RIAs does not take into consideration that the contract labs are reporting performance of their assay that falls outside the performance standards reported by the manufacturer.

CHARGE QUESTIONS

1. CLARITY OF THE STATED PURPOSE OF THE ASSAY.

The purpose of the assay is clear. It is difficult to imagine what a novice in this field would require to perform the assay as intended by the EPA; presumable, the contract labs performing this would have experience.

2. CLARITY, COMPREHENSIVENESS AND CONSISTENCY OF THE DATA INTERPRETATION WITH THE STATED PURPOSE OF THE ASSAY.

A significant weakness is that the interpretation of the thyroid endpoints seems to require previous knowledge of the activity of the test compound. The justification of this statement is that the profile of effects observed with Phenobarbital was identical to that observed with other compounds like Linuron and Flutamide, yet the interpretation was based on previous

publications. Moreover, all hormone levels are not being measured in a manner that will lead to logical interpretation.

3. BIOLOGICAL AND TOXICOLOGICAL RELEVANCE OF THE ASSAY AS RELATED TO ITS STATED PURPOSE.

The assay is relevant to the goals of the EDSP. Data from this assay will not likely provide novel biological information, although it could provide the motivation to address specific mechanistic hypotheses.

4. CLARITY AND CONCISENESS OF THE PROTOCOL IN DESCRIBING THE METHODOLOGY OF THE ASAY SUCH THAT THE LABORATORY CAN:

A. COMPREHEND THE OBJECTIVE.

The objectives are clearly and concisely articulated in the protocol.

B. CONDUCT THE ASSAY

To the best of my ability to determine, the protocol is sufficiently detailed that an experienced laboratory could conduct the assay as written.

C. OBSERVE AND MEASURE PRESCRIBED ENDPOINTS

Generally, the prescribed endpoints are clearly articulated. In addition, the EPA has developed performance standards, which will likely improve the quality of the data received from this assay. However, the mechanism by which these performance standards are generated should be more closely evaluated. Moreover, there are no performance standards established for the RIAs. These commercial RIA kits come with manufacturer-established performance characteristics, but the EPA does not require that contract labs use these kits in a manner that is consistent with the manufacturer performance. Finally, some of the kits being used in this assay are heterologous (i.e., prepared and calibrated for human samples, but used in rats), and the EPA does not require the contract lab to validate the assay. Given this situation, it is no wonder that there is a high degree of variation in hormone measurements. It is highly likely that the variability in hormone levels observed in these experiments can be reduced to such an extent that hormone levels themselves can play a larger role in the in vivo portion of the EDSP.

D. COMPILE AND PREPARE DATA FOR ANALYSES

The protocol is clear in directing laboratories in their data preparation and analysis.

E. REPORT RESULTS.

Likewise, the kind of information requested in the report from studies are clearly represented in the protocol.

5. STRENGTHS AND/OR LIMITATIONS OF THE ASSAY IN THE CONTEXT OF A POTENTIAL BATTERY OF ASSAYS TO DETERMINE INTERACTION WITH THE ENDOCRINE SYSTEM.

The strength of the assay is that it evaluates the interaction of toxicants with the androgen and thyroid systems during a particularly sensitive period of development for these interactions. Thus, this assay should be more sensitive that the adult in identifying EDCs. The endpoints for androgen action are valid and known to be sensitive to changes in hormone action during this period. There are no endpoints of thyroid hormone action that are equivalent to those of the androgen system (e.g., secondary sex organ weight). This produces an unbalanced assay that has confounded the interpretation of results of this assay and the Adult Intact Male 15-day Assay. Specifically, several compounds cause a reduction in serum total T_4, including Linuron, PBDEs, PCBs, Phenobarbital. In each case, the interpretation is that these compounds activate liver enzymes (e.g., UDPGTs) that decrease circulating half-life for serum T_4. However, in the case of Phenobarbital (mostly), serum TSH is increased in response to low T_4. In contrast, serum TSH is not always elevated in response to low T_4 produced by Linuron, or PCBs. It is not clear why the same level of serum T_4 is not always associated with an increase in T_4, but this apparently is the case. The question is whether low serum T_4 produces adverse effects on peripheral tissues – especially during development – in the absence of increased TSH. Failure to identify and incorporate valid endpoints of TH action that would be equivalent to (e.g.,) seminal vesicle weight for androgens, represents a significant weakness in the EDSP that will create considerable debate about the interpretation of data derived from these assays.

6. IMPACTS OF THE CHOICE OF: A) TEST SUBSTANCES, B) ANALYTICAL METHODS, AND C) STATISTICAL METHODS IN TERMS OF DEMONSTRATING THE PERFORMANCE OF THE ASSAY.

The test substances were appropriately chosen as chemicals that produced a relatively weak effect. This was a good test of the sensitivity of the assay and resulted in somewhat ambiguous results with Phenobarbital. The analytical methods were variable. Certainly body and organ weights are reasonably analyzed. However, histopathological analysis should be objectified with computer-assisted morphometry. Morphometry has been shown to be more reliable than visual estimation (1). In addition, the RIAs are problematic as has been discussed above. The description of the statistical methods is clear and appropriate.

7. REPEATABILITY AND REPRODUCIBILITY OF THE RESULTS OBTAINED WITH THE ASSAY, CONSIDERING THE VARIABILITY IN THE BIOLOGICAL AND CHEMICAL TEST METHODS.

An important question is whether the variability in endpoints can be reasonably reduced – both within lab and between labs – by standardizing different elements of the test. A major variable will be that of the feed. We know that variation in the amount of estrogenic compounds in feed is high, regardless of the supplier's certification. This variation alone can interact with test compounds to provide variability from lab to lab, or at different times within the same lab. Variability in hormone levels will be affected by this, but also by the standardization methods as described above.

1. **Hooth MJ, Deangelo AB, George MH, Gaillard ET, Travlos GS, Boorman GA, Wolf DC** 2001 Subchronic sodium chlorate exposure in drinking water results in a concentration-dependent increase in rat thyroid follicular cell hyperplasia. Toxicol Pathol 29:250-259

Appendix A

CHARGE TO PEER REVIEWERS

PEER REVIEW CHARGES

for

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

September 13, 2007

Background:

According to Section 408(p) of the EPA's Federal Food Drug and Cosmetic Act, the purpose of the EDSP is to:

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

Subsequent to passage of the Act, the EPA formed the Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC), a panel of scientists and stakeholders that was charged by the EPA to provide recommendations on how to implement the EDSP. Upon recommendations from the EDSTAC, the EPA expanded the EDSP using the Administrator's discretionary authority to include the androgen and thyroid hormone systems as well as wildlife.

One of the test systems recommended by the EDSTAC was the male pubertal rat assay. The purpose of the pubertal assay is to provide information obtained from an *in vivo* mammalian system that will be useful in assessing the potential of a chemical substance or mixture to interact with the male endocrine system. This assay is capable of detecting chemicals with antithyroid, androgenic, or antiandrogenic [androgen receptor (AR) or steroid-enzyme-mediated] activity or agents which alter pubertal development via changes in gonadotropins, prolactin, or hypothalamic function.

Briefly, the study design uses weanling rats, standardized to 8 - 10 per litter at post-natal day (PND) 3-5, that are housed 2 to 3 per cage. The test chemical is administered in corn oil by oral gavage (2.5 to 5.0 ml/kg) between 0700 and 0900 (lights 14:10, on 0500h) from PND 23 - 53 (31 days) to 15 males per dose level. The endpoints are growth (body weight); age at preputial separation; serum testosterone, thyroxine (T_4) and TSH; weights of reproductive organs (seminal vesicle plus coagulating gland (with and without fluid), ventral prostate, dorsolateral prostate, levator ani plus bulbocavernosus muscle complex, epididymis, testis); histology of epididymis, testis, thyroid, and kidney; and weights of thyroid, liver, kidney, adrenal, and pituitary.

Although peer review of the male pubertal 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 expected to be used among assays within the Tier-1 battery to determine whether a chemical substance interacts with the endocrine system. 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).

Each peer reviewer is asked to review the Integrated Summary Report and protocol (Appendix 1), and comment on the results of the validation process of the male pubertal rat assay, especially the inter-laboratory validation exercise[1]. Review and comment shall be directed to each of the following:

1. Clarity of the stated purpose of the assay.

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

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

4. Clarity and conciseness of the 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. Strengths and/or limitations of the assay in the context of a potential battery of assays to determine interaction with the endocrine system.

6. Impacts of the choice of:
    a. test substances,
    b. analytical methods, and
    c. statistical methods in terms of demonstrating the performance of the assay.

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

Appendix B

INTEGRATED SUMMARY REPORT

Integrated Summary Report for the Validation of the Male Pubertal Rat Assay as a Potential Screen in the Endocrine Disruptor Screening Program Tier-1 Battery (PDF) (128pp, 550K)

Appendix C

SUPPORTING MATERIALS

Appendix 1. Male Pubertal Protocol

Protocol for the Male Pubertal Rat Assay (PDF) (19pp, 111K)

Explanation of the 5 Thyroid Slides (attached below) and How they are Used in the Protocol for the Male Pubertal Assay (PDF) (1pp, 8K)

Image of Slide 30576F1C5 (see above) (PDF) (1pp, 374K)

Image of Slide 30590F3C3 (see above) (PDF) (1pp, 424K)

Image of Slide 30593F4C2 (see above) (PDF) (1pp, 420K)

Image of Slide 30594F5C1 (see above) (PDF) (1pp, 434K)

Image of Slide 30648F2C4 (see above) (PDF) (1pp, 413K)

Appendix 2. Detailed Review Paper

The detailed review paper for the Male Pubertal Rat Assay, Endocrine-disrupting chemicals: prepubertal exposures and effects on sexual maturation and thyroid function in the male rat. A focus on the EDSTAC recommendations is copyrighted material and cannot be disseminated.

If you wish to locate a copy of the paper, please use the following citation:

Stoker, T. E., Parks, L. G., Gray, L. E., and Cooper, R. L. (2000). Endocrine-disrupting chemicals: prepubertal exposures and effects on sexual maturation and thyroid function in the male rat. A focus on the EDSTAC recommendations. Crit Rev.Toxicol. 30, 197-252.

Appendix 3. Literature Studies since the Detailed Review Paper

Table of Literature Studies since the Detailed Review Paper (PDF) (5pp, 75K)

List of References from which Data were Extracted by Contractor for the Male Pubertal Assay: Pubertal Assays since 2000 (PDF) (2pp, 5K)

Extracted Data

Ashby and Lefevre (2000) (XLS) (12pp, 340K)

Blystone et al 2007 (XLS) (12pp, 75K)

Grote et al 2004 (XLS)(10pp, 57K)

Marty et al 2001a (XLS) (12pp, 67K)

Marty et al 2001b (XLS) (12pp, 75K)

Romualdo et al 2002 (XLS) (11pp, 34K)

Shin et al 2002 (XLS) (11pp, 40K)

Stoker et al 2000 (XLS) (10pp, 49K)

Stoker et al 2002 (XLS) (10pp, 86K)

Stoker et al 2004 (XLS) (10pp, 65K)

Stoker et al 2005 (XLS) (10pp, 35K)

Stoker et al 2006 (XLS) (11pp, 45K)

Tan et al 2003 (XLS) (11pp, 40K)

Yamasaki et al 2002 (XLS) (11pp, 37K)

Appendix 4. Transferability Study (TherImmune 1) Summary Report

Summary Report of Transferability Study (Block 1) - TherImmune Research Corporation: Assessment of Pubertal Development and Thyroid Function in Juvenile Male Rats (PDF) (190pp, 5M)

Summary Report of Transferability Study (Block 2) - TherImmune Research Corporation: Assessment of Pubertal Development and Thyroid Function in Juvenile Male Rats (PDF) (198pp, 5.4M)

Appendix 5. Transferability Study (TherImmune 1) Detailed Table of Results

Detailed Table of Results from Transferability Study - TherImmune Research Corporation: Assessment of Pubertal Development and Thyroid Function in Juvenile Male Rats (XLS) (39pp, 239K)

Appendix 6. Multi-Chemical Study (RTI) Summary Report

Summary Report from the RTI Multi-Chemical Study - Assessment of Pubertal Development and Thyroid Function in Juvenile Male CD® (Sprague-Dawley) Rats After Exposure to Selected Chemicals Administered by Gavage on Postnatal Days 23 to 52/53 (PDF) (972pp, 16.1M)

Final Pathology Report from the RTI Multi-Chemical Study - Assessment of Pubertal Development and Thyroid Function in Juvenile Male CD® (Sprague-Dawley) Rats After Exposure to Selected Chemicals Administered by Gavage on Postnatal Days 23 to 52/53 (PDF) (62pp, 5.2M)

Feed Analysis Reports from the RTI Multi-Chemical Study - Assessment of Pubertal Development and Thyroid Function in Juvenile Male CD® (Sprague-Dawley) Rats After Exposure to Selected Chemicals Administered by Gavage on Postnatal Days 23 to 52/53 (PDF) (8pp, 388K)

Caveat Regarding the Analysis of the Covariance in Organ Weights in the RTI Multi-Chemical Study - Assessment of Pubertal Development and Thyroid Function in Juvenile Male CD® (Sprague-Dawley) Rats After Exposure to Selected Chemicals Administered by Gavage on Postnatal Days 23 to 52/53 (PDF) (1pp, 8K)

Appendix 7. Multi-Chemical Study (RTI) Detailed Table of Results

Detailed Table of Results from the RTI Multi-Chemical Study - Assessment of Pubertal Development and Thyroid Function in Juvenile Male CD® (Sprague-Dawley) Rats After Exposure to Selected Chemicals Administered by Gavage on Postnatal Days 23 to 52/53 (XLS) (13pp, 121K)

Appendix 8. Multi-Chemical Study (RTI) ANCOVA with Body Weight at Weaning

Analysis of Covariance (ANCOVA) with Body Weight at Weaning of Data from the RTI Multi-Chemical Study Report - Assessment of Pubertal Development and Thyroid Function in Juvenile Male CD® (Sprague-Dawley) Rats After Exposure to Selected Chemicals Administered by Gavage on Postnatal Days 23 to 52/53 (PDF) (274pp, 353K)

Appendix 9. Multi-Dose Study (TherImmune 2) Summary Report

Final Report from TherImmune Research Corporation: Pubertal Toxicity Study of Vinclozolin, Flutamide and Phenobarbital in Male Sprague Dawley Rats and Methoxychlor, Ethinyl Estradiol and Phenobarbital in Female Sprague Dawley Rats when Administered in Corn Oil by Oral Gavage (PDF) (844pp, 14.7M)

Caveat regarding the Analysis of the Covariance in Terminal Body Weights in the TherImmune Research Corporation: Pubertal Toxicity Study of Vinclozolin, Flutamide and Phenobarbital in Male Sprague Dawley Rats and Methoxychlor, Ethinyl Estradiol and Phenobarbital in Female Sprague Dawley Rats when Administered in Corn Oil by Oral Gavage (PDF) (1pp, 8K)

Appendix 10. Multi-Dose Study (TherImmune 2) Detailed Table of Results

Detailed Table of Results from TherImmune Research Corporation Report Study: Pubertal Toxicity Study of Vinclozolin, Flutamide and Phenobarbital in Male Sprague Dawley Rats and Methoxychlor, Ethinyl Estradiol and Phenobarbital in Female Sprague Dawley Rats when Administered in Corn Oil by Oral Gavage (XLS) (13pp, 121K)

Appendix 11. Multi-Dose Study (TherImmune 2) ANCOVA with Body Weight at Weaning

Analysis of Covariance (ANCOVA) with Body Weight at Weaning of Data from the TherImmune Research Corporation Report Study: Pubertal Toxicity Study of

Vinclozolin, Flutamide and Phenobarbital in Male Sprague Dawley Rats and Methoxychlor, Ethinyl Estradiol and Phenobarbital in Female Sprague Dawley Rats when Administered in Corn Oil by Oral Gavage (PDF) (72pp, 121K)

Appendix 12. White Paper on Rat Strain Differences

White Paper on Species/Strain/Stock in Endocrine Disruptor Assays (PDF) (97pp, 563K)

Appendix 13. Reviewer's Comments on White Paper on Rat Strain Differences

Reviewer's Appendix to the White Paper on Species/Stock/Strain in Endocrine Disruptor Assays (PDF) (97pp, 546K)

Appendix 14. Interlaboratory Validation Study Summary Report (Charles River/Argus)

Summary Report for the CR-DDS Argus Division's Interlaboratory Validation of the Male Pubertal Assay (PDF) (48pp, 311K)

Appendix 15. Interlaboratory Validation Study Summary Report (Huntingdon)

Summary Report for the Huntingdon Life Science's Interlaboratory Validation of the Male Pubertal Assay (PDF) (49pp, 1.9M)

Appendix 16. Interlaboratory Validation Study Summary Report (WIL)

Summary Report for the WIL Research Laboratories' Interlaboratory Validation of the Male Pubertal Assay (PDF) (59pp, 651K)

Appendix 17. Interlaboratory Validation Study Analysis Report (Battelle)

Battelle Report on the Analysis of the Interlaboratory Validation of the Male Pubertal Assay: Assessment of Pubertal Development and Thyroid Function in Juvenile Male Rats (PDF) (127pp, 13.5M)

Appendix 18. Interlaboratory Validation Study Detailed Table of Results

Detailed Table of Results Extracted from the Battelle Report on the Analysis of the Interlaboratory Validation of the Male Pubertal Assay: Assessment of Pubertal Development and Thyroid Function in Juvenile Male Rats (PDF) (18pp, 276K)

Appendix 19. Interlaboratory Validation Study Comparison of Results Table

Table with the Comparison of Results from the Male Pubertal Interlaboratory Validation Study (PDF) (18pp, 115K)