Peer Review Results for the Female 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

21 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 female 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 endocrine system. This assay is capable of detecting chemicals with antithyroid, estrogenic, or antiestrogenic [estrogen receptor (ER) or steroid-enzyme-mediated] activity or agents which alter pubertal development via changes in gonadotropins, prolactin, or hypothalamic function.

Although peer review of the female 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 Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) Scientific Advisory Panel (SAP).

The purpose of this peer review was to review and comment on the female 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 female rat assay will include this peer review report consisting of the peer review comments, as well as documentation indicating how peer review comments were addressed by EPA, and the final EPA work product.

1.1 Peer Review Logistics

ERG initiated the peer review for the female pubertal rat assay on October 16, 2007. ERG held a pre-briefing conference call on October 29, 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 November16, 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.

- ***Jeffrey Blaustein, Ph.D.,*** is a Professor and Division Head of Neurobehavioral Science in the Department of Psychology at the University of Massachusetts – Amherst. Dr. Blaustein was also the Founding Director of the Center for Neuroendocrine Studies. His research efforts focus on neuroendocrinology, hormone-neurotransmitter interactions, hormones and behavior, and reproductive endocrinology. His laboratory studies the cellular processes by which steroid hormones act in neurons, particularly with respect to their involvement in reproductive behavior. Dr. Blaustein is currently Editor-in-Chief of the professional journal, *Endocrinology*, and serves on the editorial board for both the *Journal of Neuroendocrinology* and *Frontiers in Neuroendocrinology*. He is a member of the Society for Neuroscience, the Endocrine Society, the International Society for Behavioral Neuroscience, as well as others. Dr. Blaustein has published numerous

articles throughout his career that have appeared in scientific journals, such as, *Endocrinology; Journal of Comparative Neurology; Journal of Neuroendocrinology; and the American Journal of Physiology*..

- ***Kenneth Delclos, Ph.D.,*** is presently a Pharmacologist at the U.S Food and Drug Administration's (FDA) National Center for Toxicological Research in Jefferson, Arkansas. He is also an Adjunct Faculty member in the Interdisciplinary Toxicology Program, Department of Pharmacology and Toxicology, at the University of Arkansas for Medical Sciences. Dr. Delclos has given talks and presentations related to endocrine disruptors and the reproductive and developmental toxicity of phytoestrogens. He is a member of the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) Endocrine Disruptor Working Group. Throughout his 25-year career, he has published research articles in *Chemical Biological Interactions, Neurotoxicological Teratology, Reproductive Toxicology and Toxicology*.

- ***J. David Furlow, Ph.D.,*** is an Associate Professor at the University of California, Davis, in the Section of Neurobiology, Physiology, and Behavior, College of Biological Sciences, where he has served on the faculty since 1998. Dr. Furlow received his bachelor's degree in Biochemistry from the Pennsylvania State University and his Ph.D. in Biochemistry at the University of Wisconsin. At Wisconsin, Dr. Furlow did his thesis work on estrogen receptor structure and function with Dr. Jack Gorski. His post-doctoral training was done at the Carnegie Institution of Washington Department of Embryology, in the laboratory of Dr. Donald Brown. It was at the Carnegie Institution where he began working on the problem of thyroid hormone control of metamorphosis in Xenopus laevis, research that is ongoing in his current laboratory at Davis. The Furlow lab investigates the control of gene expression by the thyroid hormone receptors during metamorphosis, with additional related interests in the impact of environmental chemicals in modulating thyroid hormone receptor activity and the development of synthetic thyromimetic compounds. In addition to research on thyroid hormone action, Dr. Furlow has an active collaboration with Dr. Sue Bodine at Davis on reciprocal control of skeletal muscle atrophy by corticosteroids and IGF-1 in rodent models. Dr. Furlow has served as a faculty member and director of the Physiology summer course at the Marine Biological

Laboratory in Woods Hole, MA and most recently as a faculty member at a gene expression course run by the European Molecular Biology Laboratory in Heidelberg, Germany. He has authored 26 articles on estrogen and thyroid hormone function, and has served on several ad hoc grant review committees for both the National Institutes of Health and the National Science Foundation. Research in Dr. Furlow's laboratory is currently funded by grants from the National Institutes of Health, the Muscular Dystrophy Association, and the Netherlands Organization for Scientific Research.

- ***Heather Patisaul, Ph.D.,*** is an Assistant Professor in the Department of Zoology and an Associate Member of the Department of Toxicology at North Carolina State University. She earned her Ph.D. in Population Biology, Ecology, and Evolution from Emory University in Atlanta, GA. Dr. Patisaul studies the effects of phytoestrogens on estrogen dependent processes, reproductive behavior, and neuroendocrine systems. In 2006, she served as Chair for the symposium "Dietary influences on Aggression at the Annual Meeting of the International Society for Research on Aggression in Minneapolis, MN, and in 2005 was Chair for a workshop at the annual meeting of the Society for Behavioral Neuroendocrinology in Austin, TX. Dr. Patisaul is a member of the following professional societies: Society for Neuroscience, Triangle Consortium for Reproductive Biology, and Woman in Endocrinology. She has published her research in peer-reviewed journals such as, *Behavioral Neuroscience, Brain Research, Endocrinology, and Neurotoxicology* to name a few.

- ***Deodutta Roy, Ph.D.,*** is a80 Professor and Chair in the Department of Environmental and Occupational Health at Florida International University in Miami. His active research programs focus toward understanding the involvement of natural estrogen and estrogen-like environmental and industrial chemicals in the etiology of human urogenital cancers and reproductive diseases. Dr. Roy's research interests include, gene-environment interactions, genetic polymorphism in environmentally susceptible genes, signal transduction in toxicology and environmental health, and to develop exposure-effect assessment biomarkers. He is a member of the working group for preparing WHO organized International Agency for Research on Cancer Publication of Monographs of Hormone Replacement Therapy, the U.S. Army Review Panel on Endocrinology, the

Society of Toxicology, and a permanent member of Frontiers in Bioscience Society of Scientists. He has published over 80 peer reviewed research articles in *Biochemical Pharmacology*, the *Journal of Carcinogenesis*, *Molecular Biology*, the *Journal of Steroid Biochemistry*, and the *Journal of Toxicological Health*. He has written several monographs, proceedings and book chapters, as well as presented over 100 papers at national symposiums, including the American Public Health Association's 133rd Annual Meeting and Exposition, Proceedings of the American Association of Cancer Research, the Toxicologist, and several Society of Toxicology annual meetings.

2.0 PEER REVIEW COMMENTS ORGANIZED BY CHARGE QUESTION

Peer review comments received for the female 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.

David Furlow: The female prepubertal rat assay seeks to develop a reproducible and sensitive screening assay for the influence of xenobiotics on endocrine related endpoints, within an intact animal. The assay is sensitive and reproducible for the panel of compounds tested, and the endopoints measured are appropriate as a Tier 1 screen in conjunction with a battery of other in vitro and in vivo tests. The major drawbacks include continued concern about rat strains chosen for the study, the window of exposure to the compounds, and the lack of behavior endpoints in the assay. Most of the detailed comments are in response to questions 4 and 5.

Heather Patisaul: Critique Summary. The need for an effective pubertal assay is unequivocal. There is sufficient evidence to suggest that the proposed protocol will be able to detect a large cohort of compounds that are capable of disrupting the endocrine system and thus alter the timing of puberty. Disruption of thyroid activity will also be detected in this assay although the inclusion of these endpoints complicates the design of the assay and introduces a number of caveats as discussed in detail below. Although the assay is generally well constructed, there are a number of critical issues that diminish the biological and toxicological relevance of the assay.

The assay would benefit from simplification. A few salient endpoints, collected well and with careful controls, would be far superior to the broad spread of largely unrelated endpoints currently proposed. Inclusion of both estrogen and thyroid related endpoints complicate the interpretation of the data and as such, in many cases, the results of the data are listed as “difficult to interpret.” For the assay to be effective and reliable a data set that is “difficult to interpret”

should be rare rather than the norm. A screening assay should yield straightforward results that clearly and quickly identify compounds that require further testing.

A number of elements within the protocol diminish the functional utility of the assay. Ovarian and uterine weights are generally uninformative and complicated by cycle. Inclusion of the thyroid endpoints precludes the use of a needed positive control group for the pubertal measures. The duration of estrus monitoring is too short and the use of daily lavage will likely induce pseudopregnancy in some animals, potentially confounding the data. Failure to eliminate phytoestrogens introduces an unnecessary confound and increases the risk of inter-laboratory variability. Finally, only two doses are to be used, both of which are based on body weight and neither of which will approximate a "typical" human or wildlife exposure. The failure to include a dose within a reasonably physiological range is a considerable concern.

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

Jeffery Blaustein: The purpose is clear.

Barry Delclos: The statement of purpose of the assay is reasonably clear. A revision to the initial sentence to state more specifically what will actually be accomplished with this assay might be appropriate. For example: "The purpose of this protocol is to identify chemicals that affect, after oral administration, pubertal development and thyroid function in the intact juvenile/peripubertal female rat."

One thing that is unclear is why the oral route was selected as the only possible route of exposure for this assay. While there is likely to be limited information on the majority of chemicals that will be tested in the Endocrine Disruptor Screening Program (EDSP), in cases where either planned use or pharmacokinetic data are available, it would seem that, at least in some cases, other routes might be preferred. If the goal of the assay is primarily to detect endocrine system activity that would then be further investigated and defined in higher tier studies, it would seem that a route that results in higher systemic exposure to the test chemical might be selected even if that route is not the major route of exposure to be expected in humans. At any rate, a statement

of why the oral route is specified (e.g., because it is likely to be the primary route of human exposure, etc.) would be appropriate.

David Furlow: The assay's stated purpose is to test the effect of xenobiotics on the endocrine system of the prepubertal female rat, specifically with regard to estrogenic/antiestrogenic and antithyroid activity or disrupted hypothalamic and pituitary function as relates to the onset of puberty. This appears straightforward enough; however, in the peer review charges, the effect of mixtures is mentioned as an application of the assay. No instruction of how best to perform mixture analysis is provided (an increasingly important issue in toxicology), nor is it addressed in the ISR.

Heather Patisaul: The purpose of the assay is clearly stated and well justified. The protocol is intended to detect alterations in sexual maturation and thyroid function by exogenous chemical exposures during the prepubertal period.

It should be noted that the majority of endpoints within the assay are related to puberty rather than thyroid function. The thyroid endpoints, although interesting in their own right, feel out of place in the context of the assay. They also complicate the interpretation of the results (as discussed more in depth below) and preclude the inclusion of a positive control for the estrogenic endpoints. A "female pubertal assay" should be specific, simple, and straightforward. If possible, it would be advisable to develop a separate assay to assess thyroid function.

Deodutta Roy: Though the intent of pubertal assay described in both ISR and the Appendix 1 female pubertal protocol is the same, the description of the purpose of the assay is not exactly the same. For example,

(i) In the Appendix 1 protocol, Section I. Purpose and Applicability, the first sentence states that "the purpose of this protocol is to quantify the effects of chemicals on pubertal development and thyroid function in the intact juvenile/peripubertal female rat."

On page 8, para 3 of ISR, Section III: Purpose of the assay, first sentence: It states that "The purpose of the female 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 endocrine system."

(ii) The Appendix 1 protocol in the section I. Purpose and Applicability, second sentence: It states that "this assay detects chemicals that display antithyroid, estrogenic, or antiestrogenic activity (e.g., alterations in receptor binding or steroidogenesis), or alter hypothalamic function or gonadotropin or prolactin secretion".

On page 8, para 3 of ISR, section III Purpose of the assay, second sentence: It states that "This assay is capable of detecting chemicals with antithyroid, estrogenic, or antiestrogenic activity or agents which alter pubertal development via changes in luteinizing hormone, follicle stimulating hormone, prolactin or growth hormone levels or via alterations in hypothalamic function.

Page 3 and the Table 1 recommended by the Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC) of ISR: Pubertal female (rat): "An assay to detect chemicals that act on estrogen or through the hypothalamus-pituitary gonadal (HPG) axis that controls the estrogen and androgen systems. It is also enhanced to detect chemicals that interfere with thyroid system".

(iii) On a similar smaller note, the ISR title is "Validation of a Test Method for Assessment of on Pubertal Development and Thyroid Function in Juvenile Female Rats as a Potential Screen in The Endocrine Disruptor Screening Program Tier 1 Battery". In this title the context of use of the report is very clear, i.e., validation. However, in the pubertal assay protocol of the Appendix 1, the title starts from "Test Method----." The context in which this pubertal protocol would be used was not clear in the title. Moreover, the title of both ISR and appendix 1 should be"------Intact Juvenile/Peripubertal Female Rats" instead of "----Juvenile Female Rats". This will be consistent with the description in the purpose of the assay and also removes the ambiguity of intact versus ovariectomized.

The above different descriptions in the female pubertal protocol and in the ISR report of the purpose of female pubertal rat assay clearly create ambiguity and confusion, and it does not take into the account new knowledge in the field of endocrine disruptors.

For example, patterns of gonadotropin secretion during puberty in girls have become clearer as measurement techniques have improved. It is now widely recognized that endocrine or paracrine factors different from gonadotropins may play a relevant role as modulators of estrogen (E2) secretion early in the process of ovarian maturation that leads to premature sexual development in girls. A variety of growth factors, including IGF-I, are considered to have a synergistic effect on gonadotropin-induced stimulation of ovarian steroid synthesis or aromatization and breast development.

The purpose of the assay in ISR needs to be re-worded to remove the ambiguity and to make it comprehensive and clear. Here is a draft of an attempt to re-word it: “The purpose of the female pubertal assay is to assess the potential of a chemical substance or mixture to interact with endocrine system which influences pubertal development and thyroid function in the intact juvenile/peripubertal female rat. This assay measures indices of pubertal development and is capable of detecting chemicals interacting with the estrogen, androgen, and thyroid hormonal systems, or agents which alter pubertal development via changes in luteinizing hormone, follicle stimulating hormone, prolactin or growth hormone levels or via alterations in hypothalamic function.”

(iv) Table 3 on page 9 in the Section III: Purpose of the assay lists the end points for the female pubertal protocol: It is not clear why one of the highly sensitive hormone dependent organs, i.e., mammary gland is not included for the analysis of its weights and histopathology. In the various strain of rats, it has been shown that the treatment of 14-21 days with endocrine disruptors, particularly estrogenic in nature produces profound changes in the mammary gland. The central nervous system is one of other system should have been included for analysis of its weight and histopathology, because we now that endocrine disruptors influences its development and functions. Why the levels of estrogen, androgen and progesterone were not proposed to be measured is not clear. It is the ratio of androgen and estrogen or estrogen and progesterone which determines their effects on the target organs.

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

Jeffery Blaustein: *Data interpretation:* It is unclear which data interpretation this refers to, since data interpretation is done at various stages… by the contract lab and by EPA. Having said that, clarity of the expected data interpretation may not be optimal. There a number of statistical issues to be considered in data interpretation. Many of the statements referred to, for example, an increase that was not significant. Statistically, an increase that is not significant is not an increase, and should not be phrased as though it is. Similarly, in the Summary Pubertal Interlab Results document, blue cells highlight "apparent," not statistical, dose-response relationships. Why? In biological research, all that counts is statistically significant effects. More statistical issues are discussed in section 4.d.

To be truly objective about the value of the work and to be statistically correct, the interpretation should rely on good statistical practices. The ISR in places has an appearance of wanting to "prove" the hypothesis/conclusion that this protocol is transferable.

Barry Delclos: In general, the interpretations of the data generated in the validation studies of the female pubertal protocol that are presented in the Integrated Summary Report (ISR) were clear and thorough. There were a few points that caused some confusion.

1) In the ISR discussion of the results of the validation study that included bisphenol A (ISR, page 28, lines 21-25) it is basically concluded that the assay did not detect BPA activity with the possible exception of the body weight depression. In later discussion of the effects of estrogens on body weight (ISR, page 53, paragraph starting on line 22) it appears that, based on results with ethynyl estradiol and methoxychlor, a body weight depression in the absence of an effect on other endpoints such as vaginal opening would not be interpreted as an estrogenic activity. It is not clear how the BPA data would have been interpreted for a compound that did not have the extensive body of published data that exists for BPA. If possible, a more definitive statement should be made about the sensitivity of body weight relative to the other endpoints with regard to estrogenic test compounds.

2) With regard to the discussion of BPA in the ISR, it is indicated on lines 18-19 of page 28 that body weight at vaginal opening was increased at the high dose. However, the data in Table 12 and the data presented in the laboratory study report seem to indicate that the body weight at vaginal opening was decreased at the high dose. Consideration of this result along with the ovarian weight data would likely change the conclusion reached on line 19 of page 29 that the expected estrogenicity of BPA was not detected.

3) Another instance where the ISR was in disagreement with the laboratory study report was the case of 2-chloronitrobenzene (2-CNB). The Argus report (Appendix 13) indicated that the thyroid showed histological changes consistent with a hypothyroid state, but that there were no significant changes in thyroid hormone levels. The ISR (Table 30, page 71) indicates that there was a significant decrease in T4 in the Argus study. The ISR does not mention histopathological results from the 2-CNB study thyroid component and there seems to be only sporadic use of histopathology results in interpreting the various validation studies throughout the ISR. In the case of 2-CNB, it is interesting to note that the laboratory that did not report changes in thyroid hormone changes did report treatment-related histological changes while the other laboratories reported thyroid hormone changes in the absence of histological changes. None of the laboratories reported effects on thyroid weight for this compound. One of the laboratory study reports (Appendix 15) indicated that the meaning of an isolated rise in TSH in the absence of other significant thyroid effects was unclear. The variability of hormone measurements in the thyroid assay was discussed in the ISR, but further discussion of what will constitute a positive call when results from only one laboratory are available might be helpful.

4) In the protocol (page 17), it is indicated that changes in pituitary, liver, and kidney weights should be interpreted as relevant only if there is a significant change in organ weight relative to body weights. From the data summaries and interpretation in the ISR, it was difficult to determine if this requirement was met in all cases where changes in these particular organ weights were discussed.

5) There are several instances in the ISR (for example, pages 26, 30, 31, 50, and 70) where changes in means that are not statistically significant are discussed when the changes are consistent with what was expected based on previous studies or are consistent with other

observations in the study. Comment on how such results would be used in reaching a decision on the activity of a compound might be appropriate in Section XVI (Data Interpretation) of the protocol.

6) One of the required endpoints, age at first estrus, is discussed for some of the validation studies, but not for all. For example, the discussion of the interlaboratory study does not mention how this endpoint was affected or how the result impacted the conclusions for the test compounds.

7) The detection of estrogenic activity of methoxychlor at 12.5 mg/kg in the multi-dose study (Appendix 8) is cited as an example of the sensitivity of the assay (ISR page 47, lines 14-17), but the three laboratories in the interlaboratory comparison study did not detect activity at this dose. It is of interest to note that the study that did detect activity at 12.5 mg/kg appeared to use the diet with the highest phytoestrogen level, although it is certainly not clear what factors might have contributed to the discrepant result.

David Furlow: These issues all appear appropriate; some caveats are discussed in other sections below. One item to note here is that changes in liver enzyme profile are an indicator of the presence of a potential thyrotoxicant (p. 18 Table 6), but liver enzymes are not measured in the protocol.

Heather Patisaul: In many cases, the results of the data are listed as "difficult to interpret." This is a significant concern as an EDSP Tier 1 Screening Assay should generate a data set that is relatively simple to interpret and reliably identifies compounds that require further testing. A laboratory running this assay should easily be able to conclude that a compound either produces or fails to produce an effect. The interpretation of the results should be as unequivocal as possible. The results of the study using 2-chloronitrobenzene (beginning on page 70 of the Integrated Summary Report) best illustrates how results from a compound, for which little about potential endocrine activity is known, might be interpreted. Two of the three laboratories reported significantly delayed vaginal opening. Significant changes in weight at vaginal opening, liver weight, adrenal weight and uterine weight were also observed in at least two of the laboratories. This was an unexpected finding but the authors ultimately conclude (albeit

tenuously) that 2-chloronitrobenzene interacts with the endocrine system though in indeterminate mechanism. No information is given as to how this compound would then be classified. Would it move to Tier 2 screening? Would the results be questioned? How would this data be received by the EPA? This compound was selected for screening because it was hypothesized to have no effect on the endocrine system. The data do not support the hypothesis. How would that data be used? Would it be questioned?

Deodutta Roy: In the pubertal protocol (Appendix 1), Table 6 entitled "Potential changes indicative of different mode of action that may be observed in female pubertal protocol" provides very clear comprehensive summary of expected different effects that may be observed from different modes of action. Using this table, chemical substances that that exert effects via various mechanisms or different modes of interaction with the endocrine system can be identified. The description of the data interpretation is very consistent with the stated purpose of the assay. The guidelines described in the text given for data interpretation for doses level tested, explanation of negative results in the context of interaction with endocrine system, performance criteria, and evaluations of endpoints are very clear. The same is true for ISR report which describes this in the data interpretation section (page 74-77).

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

Jeffery Blaustein: If the purpose of the assay is to quantify the effects of chemicals on pubertal development and thyroid function, then the procedures should optimize the chance of success and minimize confounds that would obscure the results. This reviewer sees a number of serious problems with the protocol that present confounds.

Supply of animals for the experiment and confound of stress exposure: First and foremost, this reviewer has a major concern in the way that the animals are received. In Section IV, paragraph 2 of the protocol, it is stated that rats are "bred in-house or purchased from a supplier as "timed pregnant" dams with arrival at the laboratory on gestation day (GD) 7, 8, 9 or 10". The use of timed pregnant animals in a reproductive study, or for that matter in most studies, is contraindicated, because shipping is a stressor, and gestation is a time of vulnerability to stress for both the fetus and the mother. Therefore, this introduces a major confound in the protocol.

Depending on how and when rats are supplied (shipped "timed-pregnant" vs. bred in lab, some animals will not be exposed to a stressor, some to a major stressor. In addition, the developmental age prenatally that the rats are exposed to the stressor will vary depending upon day of pregnancy that the rats are shipped. It is likely, but should be determined if this is a confounding factor or not, that prenatal stress influences the fetus's physiology. If there are influences, which is this reviewer's expectation, then use of "timed pregnant" females should be considered unacceptable. In-house breeding is complicated and may require additional facilities that some contract laboratories have. Therefore, it may decrease the number of laboratories equipped to do the experiments. However, use of timed pregnant animals is a serious flaw in the design. If the experiments were submitted to an endocrine journal of which I am editor, they would not be accepted.

A secondary problem with use of animals derived from mothers shipped while pregnant is the possibility that the stress of shipping compromises maternal care of the F1 generation. Since quality of maternal care is a prerequisite to normal development, and suboptimal maternal care can have epigenetic effects on the offsprings' subsequent response to hormones (Weaver et al., 2004), and perhaps xenoestrogens, use of timed-pregnant rats presents a major problem. There are so many interactions between the hypothalamo-pituitary-adrenal axis and the hypothalamo-pituitary-gonadal axis that the influence of stressors on the dependent variables has to be considered in the design of the protocol.

Necropsy. Test Method, Section X. para 1. It is stated that "On the day of kills, moving the cages or otherwise stressing the animals unnecessarily should be avoided so that variations in stress-related hormone levels are minimized." Although stress-related hormones are not being assayed in this study, this statement should still be more directive. There is no need to move or clean cages on the day of euthanasia, and it should be prohibited.

General Statement about stressors: The influence of potential sources of stressors really needs serious consideration by an expert in developmental influences of stress on physiology, and not just the influence of stress on stress-related hormones.

Diet: The ISR (page 78-79) makes the argument that, although phytoestrogens in the diet may have effects on vaginal opening, this is unlikely to be a concern, because control groups will be exposed to the phytoestrogens as well. This demonstrates a lack of understanding of physiology. Just for the sake of argument, take the case of the presence of a hypothetical antiestrogen with no estrogenic effects in the feed. The antiestrogen might be expected to block the effects of an estrogenic test compound, but be without effect in the control group. This would then lead to a false negative. One can come up with all sorts of scenarios in which a compound in the feed would influence the experimentals, but not the controls. For example, a drug that has a permissive effect on action of an estrogen. It would have no effect in controls, but a very dramatic effect in the experimentals. I must respectfully disagree with the conclusion that it is "prudent to set a limit on the concentration of phytoestrogens in feed used in the pubertal assay." The cost for requiring that phytoestrogen-free or at least quite reduced would seem to be minimal contrasted with the cost of each of these studies.

Animal housing. Animals are housed in clear plastic cages. Problems with leaching of bisphenol A from some plastic cages have been documented (Howdeshell et al., 2003). According to these authors, polycarbonate and polysulfone cages leach bisphenol A, but polypropylene cages do not. Since this presents another potential confound, the use of particular plastics and methods for cleaning them should be given a great deal of thought, and the protocol should be very proscriptive in what is acceptable.

Littermate effects. Test Method, Section IV, para 4: It is stated "Avoid placing littermates in the same group." Since litter-effects can be so robust, this statement should be considerably stronger than it is. If the experiment is worth doing, then the protocol should be very clear that placing littermates in the same group is unacceptable.

Test Method, Section IV, para 6: The second statement is very ambiguous, and should be more directive. In addition, the statement about littermates states that "littermates should not be in the same group." Since animals were assigned to groups in paragraph 4, this seems misplaced and therefore a possible source of confusion.

Injection vehicle. Test Method, Section VI, para 2. Corn oil is the preferred vehicle. However, it is not stated that this must be of pharmaceutical grade. Relying on the experimenter to make judgments of clarity, sedimentation and odor, without specifying a common source or grade will likely lead to differences among laboratories. Likewise, leaving up to the lab the choice of corn oil, water or carboxymethylcellulose is a mistake. Although this reviewer is not a pharmacokinetics expert, the solvent would be likely to influence the rate of uptake into the circulatory system. In addition, more information needs to be given regarding the method of making up of the solutions. Should the solutions be warmed or not? Should they be subjected to sonication? Etc.

Barry Delclos: EPA has provided ample background and discussion on the biological and toxicological relevance of the assay and its ability to multiple mechanisms of interference with estrogen and thyroid hormone activities. The endpoints specified are amenable to a large scale screening program. Concern over a lack of a clear demonstration of assay specificity to this point is an issue that has been recognized and discussed. The controversial issues of sensitivity differences of rat strains and the potential impact of diet were discussed and EPA has provided reasoned explanations for their decision to recommend the Sprague-Dawley rat and to set an approximate limit of 300 ppm genistein-equivalents of phytoestrogens. These decisions will no doubt be reviewed as additional data become available. One issue that was not directly addressed in the Integrated Summary Report or the protocol itself was the question of whether confining testing to the Maximum Tolerated Dose and one half of the Maximum Tolerated Dose could miss important endocrine activity that would be evident at low doses. A footnote in the protocol to restate the EPA position on this issue should be considered.

David Furlow: Measurements of are deemed appropriate for measuring interference with estrogen and thyroid hormone endocrine systems. Biologically relevant endpoints for thyroid hormone such as T4 and TSH measurements and thyroid histology are appropriate, as are time to vaginal opening, parameters of the onset and length of the estrous cycle, and uterine histology for alterations in the function of the hypothalamic-pituitary- gonadal axis.

Heather Patisaul: The need for an effective pubertal assay is clear and well justified. However, there are a number of design issues that diminish the efficacy and relevance of the current protocol.

a. Lack of a Positive Control Group

The lack of a positive control is a serious concern. Within the Integrated Summary Report, this omission is justified by the argument that it is highly unlikely that a single compound that will generate a positive result for all endpoints in the assay. This problem results from the inclusion of experimental endpoints designed to address two different and largely unrelated questions. By lumping pubertal endpoints, which assess estrogen action, together with thyroid endpoints, the choice of an appropriate positive control becomes complicated. It is readily apparent that the most salient and critical goal of this assay is to identify compounds that affect puberty. As such, a positive control that reliably and consistently advances puberty should be included, regardless of whether or not any thyroid endpoints are altered. Estradiol, DES, or estradiol benzoate would all be appropriate positive controls and at least one should be used by all labs for this purpose. Any labs not observing an effect with the positive control would then immediately know that they have a problem executing the assay properly.

c. Selection of Dose

Only two doses will be used in the assay. Both are based on body weight with the second being half of the first. By basing the doses used for the assay on the maximum tolerated dose (MTD), both are likely to be quite high compared to what humans and wildlife could reasonable expect to be exposed to. Some of the most concerning findings with endocrine disrupting compounds have occurred at doses that are well within the realm of human exposure and far lower than would be chosen by the parameters of this assay (Alworth et al., 2002; Gioiosa et al., 2007; Goodman et al., 2006; Kato et al., 2003; Rubin et al., 2006; Rubin et al., 2001; vom Saal, 2006; vom Saal and Hughes, 2005). One of the recommendations in the Final Report of the Endocrine Disruptors Low-Dose Peer Review (2001, NEIHS) was to replicate and validate “low dose” studies. Although the argument for “low dose effects” is controversial, employment of a low dose in the assay is well justified and would address this issue. The data within the Integrated Summary Report illustrate the critical need for the employment of a lower dose. Atrazine, Bisphenol-A and methoxychlor all produced significant effects at substantially lower doses than

what would likely be used under the current protocol guidelines. Inclusion of a "low dose" would also increase the robustness of the assay. For example, within the Integrated Summary Report the observed effects of methoxychlor at 25 and 50 mg/kg/day are argued to confirm "the transferability of the assay and provide evidence that the assay is sensitive." (Page 27, line 7-8). The use of a "low dose" as defined by the Endocrine Disruptors Low-Dose Peer Review (2001, NEIHS) or a similar protocol would significantly strengthen the assay and help to clarify whether or not "low dose" effects are of genuine concern. If left out, this controversy will continue to linger and doubt about the safety of the test compounds will remain even after they are subjected to testing with this assay.

It should be noted that most of the data regarding low dose effects have come from animals exposed during the gestational or neonatal period. Therefore it is unclear if low dose effects would be observed when the exposure begins just prior to puberty, as proposed in the current protocol. However, there is sufficient data to warrant the inclusion of a low dose.

Finally, the use of high doses may explain why, to date, no compound has produced a negative result in this assay. The highest dose to be used is defined as a statistically significant reduction in body weight with "no clinical signs of toxicity." The acceptable "signs of toxicity" are not identified or discussed in the protocol but should be, perhaps in an appendix. In general, the use of body weight to define dose is problematic for several reasons, most of which have already been addressed previously by Goldman et al, but again highlights the need for a positive control group within this assay. It is well established that estradiol administration significantly reduces body weight. Because a decrease in body weight of 10% or more can result in the disinclusion of subjects or treatment groups, the employment of a positive control group would help clarify whether or not the MTD was reached or exceeded, and whether or not the laboratory was conducing the assay properly.

Deodutta Roy: The pubertal period is a very sensitive age for exposure to agents which alter the endocrine system. Therefore, this assay when validated should be able to detect chemical disruptors of estrogen, androgen, and thyroid action. The female pubertal rat assays can also identify compounds that alter hypothalamic-pituitary control of the gonads or thyroids.

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

David Furlow: In general, the protocol is well written, easy to follow, and instructions on data collection and reporting are clear.

Specific comments on the test method:

- Specific plastic used for caging (p.2 line 2) should be described in more detail. I found at least one article describing the leaching of potential endocrine disrupting chemicals from polycarbonate tubs used in rat or mouse housing, and this can be drastically increased after exposure to alkaline washing conditions and/or high temperature (Koehler et al. Lab Animal 32: 24-27, 2003; Everitt and Foster ILAR 45: 417-424 2004).
- Specific water bottles to be used for ad lib drinking are not described (see previous concerns re: housing and washing and care of plastics) (p.2 line 8).
- Within an experiment, juvenile rats should be obtained from either pregnant females from an in-house breeding or purchased from a supplier but not mixed in a study or when comparing between studies (p. 2 line 19). Inadvertent prior exposure to chemicals during gestation and the perinatal period, or maternal stressors, can influence later responses to these same or different chemicals to the offspring (e.g. Newbold et al. Reproductive Toxicology 23:290–296 (2007)). Therefore, the source of animals used in the studies ought to be standardized.
- Designation of the Maximum Tolerable Dose level is unclear to me (p. 4 line 2), since the protocol is presumably to be applied generally to compounds with both known and unknown general toxicity profiles. The way this is worded implies that preliminary studies are done to determine the MTD prior to a full scale assay for endocrine system effect, which may not actually be the case.
- One significant source of stress that can be avoided is to sacrifice the animals at an area of the laboratory away from the rest of the animals to be sacrificed that day. (p. 6 line 23).
- The assay protocol includes uterine wet weight as an endpoint (p. 7 line 7), yet in the ISR p. 59 line 8 it states that this is deleted from the protocol due to variability. This is unfortunate, since in ovariectomized rats, uterine wet weight is an excellent predictor of estrogenicity of a compound

although the EPA peer review web site lists a rat uterotrophic assay as one of the battery of tests to be evaluated separately.

- The lack of a standard "recommended" assay for hormone measurements (p. 7 line 28), (several options are deemed acceptable as long as quality control samples are run) may have to be revisited due to the outlying or inconsistent T4 and TSH values reported in the interlaboratory exercise (Argus laboratory values; Figure 2 page 67 and Table 30 page 73 of the ISR). This may include having a separate lab coordinate all the hormone assays, or settling on a recommended assay kit and vendor to improve reproducibility. Nevertheless, the T4 and TSH values at least changed in the same direction in all three laboratories.

Heather Patisaul: If the assay is to serve as a reliable and predictive tool for the identification of endocrine-disrupting compounds in females, it should be straightforward, clearly written, and easily performed in laboratories with reasonable testing experience. It should also be easily transferable and reproducible. There are a number of issues that impair these goals in the current protocol.

Deodutta Roy: The protocol is well written. . Examples of Tables 1-5 must be very helpful for the contract laboratories in measuring endpoints and preparing reports. It is assumed that all studies were conducted in professional contract laboratories with GLP facility. On minor points, it was not clear from the document whether each contractor purchased the animal, chemicals, and kits from the same source and the protocol of each laboratory assay was the same. This would have reduced some of the variability. The descriptive text in the protocol needs some further improvement for the clarity. As described above, the first sentence of the protocol states that "the purpose of this protocol is to quantify the effects of chemicals on pubertal development and thyroid function in the intact juvenile/peripubertal female rat." The word "quantify the effects---" is misleading because some of the data, particularly histology of thyroid section are qualititative and semi-quantitative in nature. It would be appropriate to use "determine or investigate the effects----" instead of "quantify the effects-----."

The second sentence of this section states that "this assay detects chemicals that display antithyroid, estrogenic, or antiestrogenic activity (e.g., alterations in receptor binding or steroidogenesis), or alter hypothalamic function or gonadotropin or prolactin secretion". Are these measures or effects of chemicals described in the second sentence indices of pubertal

development is not clear? The purpose of the protocol should clearly match with assay objectives with multiple endpoints. The impaired pubertal development includes early or delayed onset of puberty, impaired gonadal maturation (steroidogenesis) or ovarian function, shown by decreased or increased estrogen levels, impaired secretion of gonadotropin, thyroid hormones or prolactin. The clear connection between first two sentences is missing. The second sentence should be re-worded as "The assay is expected to identify the endocrine-mediated effects on female pubertal development by measuring puberty indices following exposure to chemicals with estrogenic or anti-estrogenic activity, inhibitors of steroid and thyroid hormone synthesis,-----."

It is not clear how much blood is needed for hormone assay and at what speed it should be centrifuged. For the methodology of hormone measurement, four methods are described and it appears that the choice to choose was left to the contract laboratory. If available, the preferred choice of assay should have been time-resolved immunofluorometric assays (IFMA) particularly for measurement of gonadotropin concentrations. This is more sensitive than radioimmunoassay or immunoradiometric assays. This would have helped in reducing intra-laboratory variability

2.5.1 Comprehend the Objective

Jeffery Blaustein: The statement of purpose and applicability is quite clear.
One small exception: Test Method, Section VII, para 1. Replace phrase "*in extremis"* by « at the point of death, » since this Latin term is not understood by all. It is also in the wrong place in the sentence, since as written, it states that they will be euthanized to the point of death, and it should probably state that the animals are found in the cage near the point of death should be euthanized.

Barry Delclos: See comment under "1. Clarity of the stated purpose of the assay" above.

Heather Patisaul: The objective is clearly stated in the protocol.

2.5.2 Conduct the Assay

Jeffery Blaustein: There are a number of problems in the description of the protocol that are likely to incorrect procedure. In general, I do not think it is sufficiently proscriptive and leaves

too much to the judgment of the experimenter. Why should any aspects of the protocol be left to the discretion of the experimenter. This would be likely to result in variability in results.

Water: It is stated that tap water in not acceptable, and deionized water is preferred. Since all labs have access to deionized water, why not simply require it to standardize the methods as much as possible? As with food, there is evidence that some water supplies can contain compounds with estrogenic properties. It would be most prudent to state requirements.

Decapitation**:** Test method, Section X. para 2. It is stated that the preferred method of kill is by decapitation without any form of anesthesia. More discussion is needed, since this statement conflicts with the statement in the Guide for the Care and Use of Laboratory Animals:

> "Euthanasia is the act of killing animals by methods that induce rapid unconsciousness and death without pain or distress. Unless a deviation is justified for scientific or medical reasons, methods should be consistent with the *1993 Report of the AVMA Panel on Euthanasia* (AVMA 1993 or later editions). In evaluating the appropriateness of methods, some of the criteria that should be considered are ability to induce loss of consciousness and death with no or only momentary pain, distress, or anxiety; reliability; nonreversibility; time required to induce unconsciousness; species and age limitations; compatibility with research objectives; and safety of and emotional effect on personnel.
>
> Euthanasia might be necessary at the end of a protocol or as a means to relieve pain or distress that cannot be alleviated by analgesics, sedatives, or other treatments. Protocols should include criteria for initiating euthanasia, such as degree of a physical or behavioral deficit or tumor size, that will enable a prompt decision to be made by the veterinarian and the investigator to ensure that the end point is humane and the objective of the protocol is achieved.
>
> Euthanasia should be carried out in a manner that avoids animal distress. In some cases, vocalization and release of pheromones occur during induction of unconsciousness. For that reason, other animals should not be present when euthanasia is performed.

> The selection of specific agents and methods for euthanasia will depend on the species involved and the objectives of the protocol. Generally, inhalant or noninhalant chemical agents (such as barbiturates, nonexplosive inhalant anesthetics, and CO_2) are preferable to physical methods (such as cervical dislocation, decapitation, and use of a penetrating captive bolt). However, scientific considerations might preclude the use of chemical agents for some protocols. All methods of euthanasia should be reviewed and approved by the IACUC.
>
> It is essential that euthanasia be performed by personnel who are skilled in methods for the species in question and that it be performed in a professional and compassionate manner. Death should be confirmed by personnel who can recognize cessation of vital signs in the species being euthanatized. Euthanatizing animals is psychologically difficult for some animal-care, veterinary, and research personnel, particularly if they are involved in performing euthanasia repetitively or if they have become emotionally attached to the animals being euthanatized. When delegating euthanasia responsibilities, supervisors should be aware of this as a p0tential problem for some employees or students."

Therefore, while decapitation would be acceptable in this case, more discussion is needed before referring to it as "The preferred method…" In addition, as discussed in the "Guide for the Care and Use…" this is not something left to untrained personnel without regard to the animals' and the technician's welfare.

***Vaginal smears*:** There are many issues in doing vaginal smears that must be considered and are not covered in the protocol (Becker et al., 2005). First and foremost, there is no discussion of how this is to be done. The technique of vaginal lavage is a bit of a craft. If the cervix is stimulated, the animal enters pseudopregnancy (aka the progestational state), an anestrous period of twice daily surges of prolactin, rescue of the corpus luteum, and elevated progesterone levels. The females do not cycle, and the vaginal smear would look like diestrous stage. Doing these by an inexperienced technician without knowledge of the problems will result in pseudopregnancy, which would confound the results of effects of xenoestrogens.

In addition, reading the slides also that takes some practice. These are not all-or-none of one cell type or another. Typically, sample photomicrographs are included in protocols to facilitate the task of the technician and to make the assessment of cell type more repeatable (Becker et al., 2005). There are also times during the light: dark cycle that result in greatest consistency, since for example, the proestrous stage of the cycle lasts only 12-14 hours (Becker et al., 2005).

The ISR (page 75) states that "regularity of cycling should be given more weight than lack of statistical significance for the difference in weight of ovary or uterus in treated animals compared to controls." It is unclear why cyclicity data seem to have been dropped, and wee not included in the summary table 30. I could not find the data or discussion of it elsewhere in the report, except in discussion of the Therimmune multi-dose study. Why was it not included, if it is in the protocol?

The ISR indicates that listing vaginal cycles as "regular" or not offered an informed summary of the data. I could not find any indication on how "regular" cycles were determined, nor what the definition was for "cycling." Although page 9 of the protocol indicates how cycle length was to be computed, it does not indicate how many cycles are needed to qualify as having cycles, nor what constitutes regular, nor how to deal with the first days, which are usually acyclic. These are complex issues. It was very surprising that daily treatment with a fairly potent estradiol did not lead to disruption of estrous cyclicity.

On page 557 of the Therimmune final report 7244-600, for example, on page 557, animal number 9186 shows ten straight days of a diestrous vaginal smear, yet on page 558 she is referred to as cyclic. She actually exemplifies acyclicity. I did not go through all of the data for examples like this, because it is clear that the instructions did not indicate precise definitions.

Minor issues with protocol

Dissection: Test Method, Section X, para 5. "The uterus is then place on a paper towel." Usually filter paper is used, since it does not stick to the wet uterus.

Test Method, Section X, para 5. "Measures to prevent drying out may be necessary if such organs cannot be weighed immediately." Protocol should state how drying out will be

prevented, and should be very directive that any drying is unacceptable. In any laboratory that I know of, these would be weighed immediately at the time of dissection. The fact that they were allowed to dry out by at least one of the contract laboratories indicates to me that these laboratories do not always fully understand standard laboratory procedure. If they are to be used, they must be told details in very specific detail.

Problems with uterine weights, etc. Test Method, Page 17, para 4. Uterine weights and ovarian weights are not meaningful. The changes over the estrous cycle are likely to outweigh any effects of treatment. Likewise body weight has the same problem. An alternative would be to have parallel groups that are euthanized prior to puberty, so that the effects of the xenoestrogens on these variables can be assessed in the absence of ovarian hormones.

Test Method, Section X, para 10. If animal is not cycling, uterus and ovary will be either heavy or low, depending on stage (actually depending on the steroid hormone profile of the animal in the preceding day, so this is not meaningful. All that is important is age at vaginal opening, and perhaps whether the animals are cycling.

Test Method, Section XVI., para 2. It is stated that "body weight loss that does not exceed approximately 10 % is an indication that MTD was approached but not exceeded." It cannot be over-emphasized that since reduction in body weight, and in particular body fat, is an outcome of estradiol treatment, body weight loss is not useful as an indicator of MTD.

Test Method, Page 17, para 5. The term "estrus" is used as in "age at first estrus"; because this is ambiguous, and often refers to behavioral estrus, it should state "vaginal estrus."

Barry Delclos: For the most part, the methodology is clearly described. I would propose the following for consideration. All pages indicated refer to the protocol.

1) In the list of endpoints under organ weights on page 1, consider indicating that the thyroid is weighed after fixation. While this is a standard practice for protocols evaluating the thyroid, highlighting this in the list of endpoints would be helpful. Also, consider indicating in Section II

of the protocol that kidney histology is optional and add clinical (serum) chemistry, blood urea nitrogen and creatinine as optional endpoints.

2) Page 2, footnote 3: What is the basis for the statement that totally synthetic diets are not appropriate? Is this because data indicate that they are not suitable, or because there are insufficient data to support their use? Most studies (for example, reference 58 in the ISR) appear to use the older AIN-76 formulations, and there appear to be less data for the revised AIN-93 diet formulation. AIN-93 has been used in reproductive studies (*e.g.* Collins *et al*., Effects of flaxseed and defatted flaxseed meal on reproduction and development in rats, Food and Chemical Toxicology **41**: 819-834, 2003). In that published study, data presented for sexual development for the female controls on the AIN-93 diet seems to fall in line with performance criteria given for the female pubertal protocol.

3) Page 2, footnote 3: The recommendation of the limit on the level of phytoestrogens is based on data from the uterotrophic assay. Recent data reported by Thigpen *et al.* http://www.ehponline.org/docs/2007/10165/abstract.html indicates that the CD Sprague-Dawley rat may be insensitive to phytoestrogens relative to the F344 rat. Response to an estrogen challenge was not reported in that study, but it would appear that there would be concern with the level of phytoestrogens specified in the current protocol if strains other than the Sprague-Dawley are used in the future.

4) Page 2, footnote 5: Is the statement that the AAALAC guideline extremes are "only marginally tolerable for the pubertal assay supported by published data that could be cited?

5) Page 3: With regard to guidance on the culling of litters, it might be useful to directly state that litters smaller than 8 can be used without adjustment.

6) Page 4, section 6: Some guidance on acceptable purity might be provided here. The methoxychlor (MXC) used in some of the validation studies had a purity of less than 90%. Earlier literature indicated that technical grade MXC was more potent than highly purified MXC due to presence of active metabolites in the technical grade material. For unknown compounds

going through the screen, will use of technical grade material be acceptable or is highly purified compound preferred? In any case, some statement might be provided.

7) Section IV: The diet that the timed pregnant females were fed by the supplier prior to shipment to the test laboratory should be reported.

8) Page 4, second paragraph: Guidance is given in this paragraph for setting the low dose to be used in the assay. For the interlaboratory validation study with 2-CDNB, it was not explained why the low dose selected did not follow this guidance.

9) Page 6, section X, Necropsy: The statements here in footnote 8 and on page 7 regarding carbon dioxide asphyxiation as inhumane seem to be at odds with current AVMA Panel recommendations http://www.avma.org/issues/animal_welfare/euthanasia.pdf . Those recommendations indicate that carbon dioxide is an acceptable method of euthanasia while decapitation is conditionally acceptable with scientific justification. Decapitation can be scientifically justified here, and the comments on carbon dioxide asphyxiation seem unnecessary.

Heather Patisaul: There are a few places within the methods where clarification is needed.

1. DIET: The diet should be free of phytoestrogens. As written, diets containing up to 300 µg/kg are allowed. The presence of phytoestrogens, even in such small quantities, needlessly impairs the sensitivity of the assay and introduces inter-laboratory variability. A number of phytoestrogen-free diets are now readily available from all of the major lab diet manufacturers so there is no reason not to exclude this potentially problematic source of endocrine disrupting compounds.

2. TRANSPORT: Animals can either be shipped timed pregnant or born in-house. This introduces gestational stress as a source of variability. Transport stress during pregnancy can delay parturition by 24-48 hours, reduce maternal care and expose the developing pups to stress hormones during gestation. All of this can affect the growth and overall well-being of the offspring. Maternal stress should be minimized as much as possible. As such, all females

should be impregnated in-house with care to avoid any unnecessary disturbance. Cage changes and all other intrusions should be minimized as much as possible.

3. DEFINING THE DAY OF BIRTH: The day of birth is defined as the "morning" of PND 0. This definition is too broad. Under this definition, an animal born at 12:10 pm on a Monday could be listed as PND 0 on Tuesday and cross fostered with animals born at 11:45 am on Tuesday. This means litters could contain animals born 18-23 hours apart. This is a problem in a protocol where many of the data interpretations are to be based on body weight. A specific time frame for the definition of PND 0, and some discussion is needed as to how animals born close to that window will be dealt with.

5. DEFINING ESTRUS: Daily lavage frequently induces pseudopregnancy in rats and can skew data regarding regularity of the estrus cycle and cycle length (Marcondes et al., 2002; Yener et al., 2007). Pseudopregnancy lasts 1-2 weeks, during which time the vaginal smear will have the appearance of persistent diestrus. No discussion as to how to deal with this frequent phenomenon is given in the assay.

6. LENGTH OF ESTRUS OBSERVATION: Animals are to be killed on PND 42. This means that estrus cyclicity will be assessed for less than two weeks and animals will be killed before they have completed the transition to full adulthood. This is an insufficient amount of time to evaluate the regularity of the estrus cycle and will likely fail to detect a number of compounds that ultimately impact endocrine action in females. The impact of endocrine disrupting compounds on the estrus cycle can be delayed by several weeks (Gallo et al., 1999; Patisaul and Polston, 2007; Rubin et al., 2001; Whitten et al., 1993). Observation of the estrus cycle for at least six weeks post puberty is strongly recommended.

7. EUTHANASIA: The use of decapitation without anesthesia is inappropriate and unnecessary. The Office of Laboratory and Animal Welfare (OLAW) within the National Institutes of Health (NIH) does not generally support the use of decapitation without anesthesia unless there are extenuating circumstances that require it. Use of CO2 asphyxiation will not alter any of the outcomes in the protocol and should therefore be used.

8. BLOOD COLLECTION: It is unclear how the blood is to be collected and prepared. The assay states that trunk blood should be collected by "inversion over a funnel" but the type of collection tube is not given. The choice of assay to measure plasma hormone levels will influence the type of collection tube to be used. In some cases, a siliconized or EDTA lined tube is required. This should be specified in the protocol.

9. HISTOLOGY: A citation for the grading scale associated with the thyroid sections is given, but no citations are listed for the evaluation of follicular development or uterine histology. Collecting uterine and ovarian weight has consistently proven to be problematic and it is not clear how these endpoints are informative. Time of vaginal opening, time of first estrus, and regularity of the estrus cycle 10-12 weeks post puberty are far more valid and reliable endpoints of pubertal alteration and are sufficient to draw conclusions about whether or not a compound should undergo Tier 2 testing. Therefore the organ weights are a time consuming and problematic component that are not needed and could be eliminated. Uterine and ovarian weights are also confounded by cycle. Although the protocol states that the estrous cycle at the time of necropsy should be taken into account, guidance as to how to do this is insufficient.

2.5.3 Observe and Measure Prescribed Endpoints

Jeffery Blaustein: As stated in section 4b, there are a number of problems that are not discussed that could lead to problems in measuring the endpoints prescribed in the protocol.

Barry Delclos: 1) Page 8, Section XII Histology: Add "(optional)" after "kidney".

2) Page 8, third paragraph in Section XII: There is mention of reporting changes in numbers of primary and atretic follicles. Is the ovarian evaluation intended to be a qualitative evaluation of a single section, or are step sections necessary?

Heather Patisaul: The data sheet for observations and measurements is sufficient and specific details about reservations regarding data collection were given above.

2.5.4 Compile and Prepare Data for Statistical Analyses

Jeffery Blaustein: ***Statistical issues.*** Statements like (ISR, page 23, line 8) "The ovarian weight was reduced, although this effect was present at p=.06, marginally beyond the p<0.05 cut-off for statistical significance" are inappropriate. Either a result is statistically significant (< .05), or it is not. Results that are not statistically significant cannot be used as support. Likewise, statements like "There was a nonsignificant increase in uterine weight (p = 0.08 adjusted for weaning), and a non-significant decrease in pituitary weight (P = 0.10)" (ISR, page 26, lin 16) are inappropriate. If the result is not statistically significant, and it is therefore due to chance, it really does not matter if it is an increase or a decrease; stating as the report does, that the non-significant increase is consistent with another paper's non-significant increase is inappropriate.

The ISR (page 68) states in reference to the methoxychlor data that "…all three laboratories did identify a similar pattern of response for this weak estrogen and this response was positive for interaction with the endocrine system at the same dose level." While this was true for vaginal opening, it was not true for ovaries, pituitary, liver, adrenal, but most importantly, the prototypical estrogen-dependent tissue, the uterus. As stated elsewhere, this indicates that either the protocol cannot be followed reliably or more likely that many of the parameters do not show reliable responses in animals treated in the way that they are in this protocol (e.g., offspring of mothers shipped during mid-gestation, ovaries intact so possibly cycling, tissues taken at a time that they are still potentially responding directly to the compounds).

The ISR (page 73) indicates that "for the chemical which had never been tested before, results were consistent across all three laboratories. This reviewer does not see data to support this statement. Just examining yellow highlights on table 30, one can see that, while this is reasonably true for vaginal opening and liver weight, there were mismatches in ovarian weight, pituitary weight, uterine dry weight and adrenal weight. This reviewer's interpretation is that either the end-points measured are not meaningful in tests of these compounds, or the protocol cannot be reliably followed with sufficient sensitivity of outcome measurements to be meaningful.

Adjustment for weight at weaning. The basis for adjusting organ weights at termination of the study for covariance with body weight at weaning is something that this reviewer does not understand. Since the groups are matched for body weight at weaning, this should relatively be a constant. Furthermore, to make this adjustment, one would have to know that weight at weaning is highly predictive of body weight at 42 days of age in untreated rats.

Barry Delclos: 1) Page 3, last sentence of the second paragraph: It is indicated that placement of litter mates in the same group should be avoided, but the last sentence in the fifth paragraph suggests that there may be situations where this could occur. It seems that either a stronger directive should be given (i.e., Do not place litter mates in the same group) or guidance on how litter mates in the same group should be reported and handled in the statistical analysis provided.

Heather Patisaul: Section XII of the protocol states that uterus, thyroid, ovary and kidney are to be evaluated for pathologic abnormalities but no guidance is given as to how the histological findings are to be quantitatively assessed. It is unclear how the histological data will contribute to the interpretation of the data.

2.5.5 Report Results

Jeffery Blaustein: In order to facilitate an understanding of the effects, standardized bar graphs should be required with a specific format. Generally, control group with mean +/- some indicator of variance, followed by the same for each of the treatments. Standard indicators of significance as would be found in a journal article should be placed above and below the bars. Although the TherImmune report had bar graphs, they were poorly set up. RTI had only very complex tables to wade through. In addition, it is standard to convert values like 0.0087 grams to 8.7 milligrams to facilitate digestion of the numbers

Barry Delclos: 1) Page 5, Section VIII, Vaginal Opening: It is indicated that documentation of vaginal threads or the appearance of pin holes are important and that it is critical that the "initiation" of vaginal opening be recorded. However, guidance on the interpretation and reporting of these endpoints is not provided.

2) Page 12, second paragraph: This is somewhat confusing given the guidance on interpreting pituitary, liver, and kidney weights given on page 17 of the protocol. There it is indicated that organ weight changes for these organs should be considered only if they change significantly relative to terminal body weight. As mentioned previously, it wasn't clear in the ISR or lab study report whether the ratios were being taken into account in the discussion of changes in the weights of these organs. It seems that for these three organs, the ratios to terminal body weight should also be reported in the table.

3) Page 17, last paragraph: I don't believe that this paragraph is necessary. Presumably, the fact that an isolated endpoint showed significance at the low dose, but not at the high dose would be taken into account in reaching a conclusion about the substance based on the results of this assay and when considered with results obtained in the other battery assays.

4) Page 18, Table 6: Consider adding a column describing anticipated changes for an estrogen antagonist.

Heather Patisaul: Data reporting is thorough and acceptable with the exception of the histological data.

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

Jeffery Blaustein: In this reviewer's opinion, this assay will add very little to a battery. For the cost involved, I would think that more direct tests of, for example, estrogenicity could be used. For example, *a uterotropic assay in ovariectomized or prepubertal rats or mice eliminates many of the confounds described in this review*.

Barry Delclos: EPA has provided a good discussion of the strengths and limitations of this assay. Strengths include the coverage of multiple mechanisms of action that are not covered in other proposed assays for the battery and exposure during a sensitive time period, although it could be argued that the perinatal period would be more sensitive for some endpoints. A major current limitation, as pointed out by the EPA, is the lack of demonstrated specificity of the assay. Another limitation is the impracticality of conducting the screen in more than a single strain to provide confidence in studies where endocrine activity is not detected.

David Furlow: *Strengths* of the assay: The major strength of the assay is that it is able to detect the activity of chemicals with regard to the endocrine system in an *intact* mammalian system. Cell culture and *in vitro* biochemical screening assays are useful to provide information on what effect a chemical *can* have on the function of the endocrine system but does not account for cell specific uptake, metabolism, effect of other circulating growth factors/hormones etc. On the other hand, *in vitro* studies provide more mechanistic information than are inferred from the *in vivo* system. Therefore as a component of a battery of assays, the prepubertal female rat assay appears relatively sensitive (see below for caveat) and standardizable to detect interactions between xenobiotic chemicals and estrogen and thyroid hormone physiology. It is especially noteworthy that a so-called "weak" estrogen, methoxychlor, and a "weak" thyroid disrupting compound DE-17 produced quantifiable and generally reproducible effects on the expected endpoints.

Limitations of the assay:

a. *Rat strain choice for these assays is still very much open for debate*. While it is apparent that the EPA is aware of potential strain influences in sensitivity to endocrine disrupting chemicals, the assay only proposes to use Sprague Dawley rat (Crl:CD(SD)). This choice is based largely on comparing only Wistar and Sprague-Dawley rats, and experience with the male prepubertal assay. However, there are several publications demonstrating increased sensitivity of Fisher F344 rats to estrogens and BPA relative to at least SD rats in multiple endpoints as pointed out in Appendix 12, several publications in the open literature, and as noted by the ISR. This point should not be dismissed in considering the utility of the assay based only on practical considerations.

b. *Exposure duration is confined to PND 22-PND 42*. The effect of fetal or perinatal exposure to chemicals is not accounted for in the current protocol. There is growing evidence that early exposure to chemicals or maternal stress can "program" adult physiology of the offspring (e.g. Newbold et al. Reproductive Toxicology 23:290–296 (2007)), in addition to sexually dimorphic behaviors (see c).

c. *The assay does not account for behavioral effects*. In many cases, low doses of estrogenic compounds can affect sexually dimorphic brain organization, at lower or similar doses to those

required to elicit morphological changes in reproductive tissues. Therefore, I was somewhat surprised that behavioral endpoints were not included in the assay, such as lordosis, for example.
d. *The specificity of the assay remains a question due to the effect of 2-CNB.* A potential negative control, 2-CNB, delayed vaginal opening and increased TSH levels. This issue is discussed in more detail below, but may be the result of increased metabolism of estradiol and/or thyroid hormones by the 2-CNB exposed liver. It is also interesting to note that even differences in caging can lead to significant differences in age of vaginal opening and estrous cycle parameters (Firlit and Schwartz, Biol. Reprod. 16:441-444, 1977) indicating the effect of non-specific stressors on these parameters although the exact mechanism ins not understood, to my knowledge.

Heather Patisaul: As discussed in detail above, the assay maximizes breadth at the expense of a robust experimental design to answer a specific question. Removal of the thyroid endpoints and the organ weights would allow for the inclusion of a needed and critical positive control group and make the data vastly easier to interpret. Simplification would also make the assay easier to replicate across laboratories. To assess pubertal disruption, only age at vaginal opening, day of first estrus, and regularity of the estrus cycle (at least 6 weeks post-puberty) is required. All other measures are extraneous and unnecessary. All compounds testing positive in this screening could then be advanced to a Tier-2 screening protocol. If the thyroid elements remain in the protocol, the use of a positive control for the pubertal endpoints is strongly recommended.

Deodutta Roy: The IRS report very clearly describes various studies of low and high doses of chemical substances conducted in the different laboratories. Each contract laboratory showed that the pubertal assays can identify compounds that alter hypothalamic-pituitary control of the gonads. All studies using thyroid-active agents showed that the female pubertal assays detect alterations in thyroid function following exposure to compounds interacting to thyroid system. I concur with EPA conclusion in regards to the strengths and weakness of various assays. The one of the major strengths of this study is that it is an in vivo assay, and it can measure the effects of both parent compounds and their metabolites. This assay estimates the interaction with the endocrine system. Additionally, this assay measures the effects of the endocrine disruptors at one of the critical time period of the development of the animal, which is highly sensitive to changes in the endocrine system. This would help in identifying weak endocrine disruptors. Use of the

redundant multiple endpoints increased the credence of the assay. This was further strengthened by the use of very well though performance criteria.

Minor weakness

It has been shown that the rodent pubertal female assay is useful for identifying potential endocrine disruptors having not only estrogenic/antiestrogenic but also androgenic/antiandrogenic activities, therefore it is not clear why androgenic/antiandrogenic activities were not monitored.

We now know the non-receptor-mediated mechanisms exist by which unknown disruptors can affect the embryo/fetus without showing positive effects on the proposed classical multiple endpoints. In situ biochemical and gene activation measurements or biomarkers for assessing pubertal development could have really helped to detect subtle changes in the endocrine systems which would be not detected otherwise by proposed multiple endpoints in this assay. ChIP on ChIP assay would have been more sensitive for screening effects of endocrine disruptors by studying changes in genes involved in androgen, estrogen, or thyroid systems.

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

David Furlow: The choice of test substances for the testing the performance of the assay are appropriate and span a spectrum of agents with distinct suspected modes of action. A few comments are included for future consideration however. For a test substance to detect thyroid hormone disruption, a more relevant choice than PTU may have been ammonium perchlorate since this compound is a known inhibitor of iodine uptake by the thyroid and is found in ground water, particularly near air force bases. Also, a more specific aromatase inhibitor such as fadrazole rather a general steroidogenesis inhibitor such as ketoconazole than may have been useful for assessing the specificity of the female rat assay in particular. Interestingly, it is apparently difficult to find a test substance known to be generally toxic but does not in some way impact any of the endocrine related endpoints in this assay. One possibility is that high doses of generally toxic chemicals induce hepatic phase I and II metabolizing enzymes and phase III

transporters as a by-product of exposure. This possibility was noted in the ISR (p 70 line 19)In this regard, screening the activity and/or mRNA expression of a panel of known steroid and thyroid hormone metabolizing enzymes in liver biopsies should be an important endpoint to consider.

The statistical analyses proposed appear appropriate, but this is not my area of expertise.

2.7.1 Test Substances

Jeffery Blaustein: ***Positive controls compounds.*** A positive control with results that are known with certainty should be used. This is essential to demonstrate that the laboratory has the expertise and laboratory conditions sufficient to support replicating a previous result. Although a high dose can be used as a secondary control, a low dose, positive control to demonstrate reliability of the laboratory should be included in the protocol.

Dose of compounds. Test Method, Section V, para 1. Two dose levels are used with the maximum tolerated dose (MTD) chosen based on a reduction of no greater than 10 % of the mean body weight for controls. Since compounds with estrogenic activity are being investigated, and since estrogens have anorectic and body weight suppressive effects, a different factor than MTD needs to be used.

In addition, these choices of doses are odd for endocrinological work. Since effects can differ at various points on the dose-response curve, usually a pilot dose-response curve is run. While the meaning of the term that is used in endocrinology… physiological levels… might not have meaning, can dosages be chosen based on reasonable expectation of level of exposure. The choice of two doses leaves a great deal to be desired, since most physiological work requires a dose-response curve.

Maximum tolerated dose: The ISR makes a number of statements such as the following (ISR, page 68): "Laboratories 2 and 3 reported that terminal body weight was significantly reduced in the high dose methoxychlor rats, both finding that this group weighed 94.1 percent of the controls. This indicates that the Maximum Tolerated Dose was reached, but that the body weight decrease compared t controls was not so severe as to interfere with endocrine endpoints." The

use of body weight data like these to draw conclusions of maximum tolerated dose are not meaningful within the context of an experiment examining the effects of a potential estrogen, since treatment with estradiol itself can suppress body weight by 20 %, and is responsible for quite dramatic shifts over the four day estrous cycle. The animals, who have lost weight, are not ill, and their body weight loss has nothing to do with maximum tolerated dose. They are not losing weight because of toxicity; this is just a normal physiological response to estradiol, so would seem to be unrelated to the toxicological term, MTD.

In discussion of specificity, the ISR mentions (page 82) that "a good faith effort was made to identify a chemical that was both toxic to other systems but without endocrine effects." It is not surprising that one could not be found, because a toxic compound will decrease body weight, and this seems to be required to demonstrate that the dose has exceeded the MTD. However, body weight loss due to toxicity would likely be accompanied by a decrease in nutrient intake. From a physiological point of view, food deprivation causes reproductive dysfunction. Unfortunately, approaching a problem like this from a toxicological view-point with little regard to the underlying endocrinology/physiology has problems. In short, any compound that compromises nutrition would be expected to have endocrine effects.

Barry Delclos: The selected test substances covered a variety of mechanisms of action purported to be detected by this assay and, for the most part, expected results were obtained. The selection of fenarimol, which apparently has a mixed mode of action that was not known at the time of selection, was described in the ISR as unfortunate, but it is likely that some unknown test compounds will interact with multiple targets and the data obtained with that compound did have utility. Inclusion of a broader range of compounds in the interlaboratory study would have provided added confidence in the performance of the assay, although it is recognized that the additional cost may have made that impractical.

Heather Patisaul: Given that no compound to date has tested negative in this assay it is difficult to evaluate the potential effects of test substance on outcome.

Deodutta Roy: The choice of test substances and analytical methods is well thought and is well described in this document. It has been identified in the report that the cost associated with

animal experiments did not allow to use appropriate ranges of positive and negative test substances, particularly weak positive controls. However, the use of weak positive compound was critical for validation of this study because most of the unknown test compounds are hormonally weak substances.

2.7.2 Analytical Methods

Jeffery Blaustein: ***Problems with uterine weights, etc.*** Test Method, P 17. Uterine weights, ovarian weights and body weights are rather meaningless in the context of this particular protocol, evidenced by many of the "acceptable ranges" given in the protocol's performance criteria. Normal physiological fluctuation over the estrous cycle in response to cyclic changes in ovarian hormones is likely to outweigh many effects of treatment. Although the protocol states that regularity of cycling should be given more weight than lack of statistical significance for the difference in weight of ovary or uterus in treated animals compared to controls, this begs the question of what a positive result is likely to be attributable to. Uterine and body weights are very informative within the context of an ovariectomized animal, but not in an ovary-intact. As pointed out in the Integrated Summary Report, "It is also important to note that the variation in the uterine weighs was expected, since uterine weights fluctuate during the estrous cycle and these females were killed on various days of their cycles."

Equally important, since animals are killed on PND 42, the last day of treatment, weights of any estrogen-responsive tissue or end-point will be a hodge-podge of direct estrogenic or antiestrogenic effects of the compound, indirect effects of the compound on the estrous cycle (Does it result in increased estradiol secretion? Does it result in long periods of anestrus?), and stage of the estrous cycle if the animals are indeed cycling. How can one know if a particular compound causes an increase in for example, uterine weight or a decrease in body weight *via* the compound, which is still available, having its uterotropic effect or body weight reducing effect, or by influencing these outcomes secondarily to perturbation of the hypothalamo-pituitary-gonadal axis (or by some other system)? There is a reason why the vast majority of endocrine studies on sex hormones is done in gonadectomized animals, and this is it. Considering the methods used, all that is important is whether the rats are cycling and age of vaginal opening, etc. But even with these variables, it must be remembered that advancing the age of vaginal opening

can occur by a direct effect of an estrogenic compound on the vagina, or it can be secondary to HPG dysregulation.

An estrogenic xenoestrogen will advance puberty, as will estradiol, but it will also increase weight of uterus, and it will decrease estrous cyclicity. Therefore, an animal can develop a high uterine weight in a number of ways, for example, by an estrogenic xenoestrogens acting directly on the uterus (*e.g.,* methoxychlor) or by lack of cyclicity with the ovary stalled in stage of follicular development (high endogenous estradiol).

Similarly, vaginal opening is an estrogen-dependent process, so this is really predominantly a bioassay for estrogenicity. There are, however, better, straightforward bioassays for estrogenicity (*e.g.*, uterotropic assays)

Barry Delclos: 1) The issue of the variability of tissue weights of the pituitary and adrenal glands has been addressed by indicating that steps need to be taken to avoid drying out prior to weighing (page 7). Weighing the organs after fixation (tissues placed in fixative immediately after removal) might also be considered as an approach to remedy the variability in these organ weights.
2) The issue of variability of hormone measurements was discussed in the ISR, and it was noted that performance criteria for TSH had not yet been established. The mean TSH levels in control animals vary widely across the various studies discussed in the ISR. The multichemical study used carbon dioxide as a method of kill (Appendix 5, page 13) and had a high control TSH level, but similar levels were seen in one of the laboratories in the interlaboratory studies where brief carbon dioxide followed by decapitation was the method of kill (Appendix 15). The two other laboratories in the interlaboratory study used the same method of kill as the latter study and both of these studies reported control levels of TSH considerably lower than those reported in Appendices 5 and 15. The Office of Research and Development (ORD) studies reported in Table 22 (page 51) of the ISR had the lowest control levels of TSH, which no doubt reflects the extensive experience of this laboratory with this assay. Perhaps strict adherence to the ORD protocol should be specified. However, while the variability of this endpoint is problematic, it appears that high control levels did not necessarily interfere with the ability of a laboratory to detect treatment effects.

Heather Patisaul: Analysis of variance is to be used with body weight at weaning as the covariate. Day of the estrus cycle is not taken into consideration in the analysis of organ weights but should be for the ovary and uterus as it likely has a greater effect on the weights of these organs than overall body weight.

Deodutta Roy: The methodologies for measuring indices of puberty are not very modern, and they may not be very sensitive in detecting the initiation and progression of molecular changes that ultimately impair the pubertal development. There is a concern that all these functional assays may not be able to detect subtle changes in the animals exposed to weak endocrine disruptors. The animal strains used are not the most sensitive to estrogenic compounds, which makes it further difficult in judging the suitability of analytical methods.

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

Jeffery Blaustein: For reasons discussed above, the statistical analysis has problems.

Barry Delclos: 1) Page 9, to paragraph: "When statistically significant effects are observed ($p < 0.05$), treatment means are examined further using appropriate pairwise comparison tests to compare the control with each dose group." This approach may be overly conservative for a screening assay for certain multiple comparison procedures. In many of the validation study procedures, Dunnett's test was used for comparisons of treatment groups to controls. The following statement is taken from Haseman *et al*., Statistical issues in the analysis of low-dose endocrine disruptor data" Toxicol. Sci. 61: 201-210, 2001: "For example, Dunnett's test is a standalone test that does not require statistical significance of an overall ANOVA to be valid. However, many investigators who used Dunnett's test required statistical significance of an overall ANOVA before making pairwise comparisons. Because the critical values for Dunnett's test were derived without consideration of an overall ANOVA, requiring this additional significance may result in a somewhat conservative test. Specifically, there were a few instances in which our reanalysis found significant pairwise differences by Dunnett's test that were not reported as significant by the study investigators who themselves also used Dunnett's test. Such

differences were apparently due to the extra requirement of a significant overall ANOVA imposed on Dunnett's test by the study investigator."

Revision of this recommendation should be considered, depending on the multiple comparisons test to be applied. While this statistical guidance may not have affected the interpretation of the validation studies, it could potentially affect conclusions in future screens.

Heather Patisaul: See above

Deodutta Roy: Statistical methodology is not my expertise, so I have decided not to make any comment regarding this.

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

Jeffery Blaustein: Repeatability is unlikely in many cases, because of the design of the experiments. There are simply too many confounds, ranging from differential exposure to stress conditions of some animals, food with potentially high levels of xenoestrogens, possible exposure to xenoestrogens from caging to tremendous variability because these are females, some of whom are cycling (and killed at random times in their very volatile cycle), some of whom are not. These have been covered extensively in previous sections.

Therimmune study issues:

It was not clear to this reviewer, if TherImmune was supposed to be following the "test method" supplied, or a completely different protocol. There were many differences, which could compromise repeatability.

In the Therimmune study, 1143-103 and 1143-101, page 12, timed pregnant rats were received on GD 12. The protocol states protocol states that the animals could be rec'd on GD 7, 8, 9, or 10. I do not know if this was a change in protocol, or lack of attention by Therimmune.

In response to my question to the EPA re: what the difference was between 1143-101 and 1143-103, we were told that the dosing period of 1143-103 was one week later than 1143-101. Since the animals were timed-pregnant and all animals were received (according to the protocols) on the same day, and the protocol called for starting the experiment on the same postnatal day, how can that be? If the dosing started one week later, then the animals would be one week older than the protocol required. If this is the case, and I have no way of determining this, it would suggest to me that the laboratory does not understand the importance of sticking to the protocol.

Although the protocol states limits of genistein allowed in the diet, Teklad 7012C was used, which is not routinely analyzed for genistein, and the experimenters did not have this analyzed. Furthermore, because phytoestrogens in the diet are a potential source of confound, the protocol could be much more directive. I see no reason that a specific diet could not be required

Page 15: It is stated that the ethynyl estradiol was dissolved in ethanol prior to dilution in corn oil. It is essential that the amount of ethanol be stated. In fact, since it is not in the protocol, it should not have been part of the procedure, because it may have resulted in the presence of ethanol in one group, and not the others.

Figure 2. Figures of means are unacceptable without an index of variance.

The protocol calls for two rats per cage; they used 3/cage, but it is possible that the protocol changed.

Barry Delclos: In general, the results reported in the validation studies, particularly the interlaboratory study conducted with three chemicals, indicate that the protocol generates data that lead to similar ultimate conclusions. Some issues that need to be clarified in interpreting the data have been raised above under responses to the earlier questions.

David Furlow: In general, repeatability and reproducibility was satisfactory, with the exception of the T4 measurements discussed under charge question 4. Furthermore, it is not unexpected that the wet weights of tissues would have so much variance and thus be of limited value since they cannot obviously be compared to wet weights prior to treatment.

Heather Patisaul: Overall the repeatability and reproducibility of the assay is good and has been sufficiently demonstrated. However, the report identifies a number of areas where the data are inconsistent. In nearly all cases, the inconsistencies are noted in endpoints that are not needed to identify the disruption of puberty. Ovarian and uterine weights are unreliable measures of endocrine disruption, particularly when collected without regard to cycle. More guidance as to how cycle should be considered when collecting organ weight is needed.

Inconsistency within the data may also result from variation in animal stress. Without sufficient controls to minimize stress it is likely that inter-laboratory variability will continue to be a problem particularly when large batches of animals have to be sacrificed on a single day and there is thus a lot of human activity in the vivarium. Great care is taken to consider variation in body weight when interpreting the data. In contrast, relatively little information is given to as to how stress was minimized or how housing conditions differed between laboratories. This is a considerable problem. Housing of other animals in the facility, particularly non-human primates, dogs, or cats can increase rodent stress substantially. Transport stress is also a concern. Pups born to dams delivered "timed pregnant" will experience more gestational stress than those born to mothers impregnated in-house. This may also affect inter-laboratory data collection and consistency. Housing conditions, including diet and day of cage changes, should be matched as closely as possible between laboratories.

Deodutta Roy: Based on the ISR document, it appears that results obtained with the pubertal assay in the different contract laboratories are repeatable and reproducible, because all three laboratories data showed that the female pubertal assay may be useful for identifying chemicals that operate through a variety of mechanisms. This was true for both estrogenic and thyroid system interacting chemicals. For example, the TherImmune 1 study used a single dose of six different compounds in three different laboratories. Three different laboratories identified expected endocrine effects from exposure to chemicals with estrogenic, anti-estrogenic, androgenic or anti-androgenic activity, inhibitors of steroid and thyroid hormone synthesis, and a dopamine antagonist. ethynyl estradiol, tamoxifen (e.g, antagonist and partial estrogen agonist), and methoxychlor advanced the onset of vaginal opening. Propylthiouracil (e.g., an inhibitor of

thyroid hormone synthesis), ketoconazole (e.g., an inhibitor of steroid synthesis) or pimozide (e.g., a dopamine antagonist) delayed the age of vaginal opening. The sensitivity of the protocol was assessed through multi-chemical study. Two different doses of six compounds were used for this study. The low doses of all six compounds showed expected changes in the estrogen-related and thyroid system-related endpoints. The multi-dose study (TherImmune1 2) used three compounds, ethynyl estradiol, methoxychlor and phenobarbital and showed similar sensitivity to estrogenic compounds. Thus, the EPA in this ISR document has very correctly concluded that the female pubertal protocol is transferable, sensitive and reproducible.

2.9 Additional Comments and Materials Submitted

Jeffery Blaustein: *Additional issues with items referred to in the Integrated Summary Report*
It is stated that they showed transferability, because all of the drugs had expected results on advancing or delaying puberty. However, tamoxifen advanced, so it is unclear why this is considered success. While it was fine for day of vaginal opening, what would be considered positive results for the other parameters?

On page 23, line 20, it is stated that large variations in weights were expected because the animals were killed on different stages of the estrous cycle. This is correct, but if so, why bother going to the trouble and expense of collecting all of these weights?

On Page 27, it is indicated that the experiment showed transferability of the protocol for methoxychlor. While it is true that the transferability of the protocol for vaginal opening was shown, they failed to find significant effects on two other parameters which would have been expected (uterine and ovary weight). My conclusion from the data would be that they did *not* replicate two previous findings.

On page 42, it is stated that ethynyl estradiol-treated rats had normal cycles. This is quite unexpected, since chronic estradiol treatment should stop the cycling by negative feedback. So, why would ethynyl estradiol result in normal estrous cycles?

Without going into great detail, my general sense from scanning the spreadsheet and Table 30 was that, although there was reasonable consistency among labs in day of vaginal opening, there was not a great deal of consistency among labs in many of the other parameters. But even in "age at vaginal opening," Lab 1 saw no effect of DE-71 or 2-chloronitrobenzene, but saw an effect of methoxychlor, Lab 2 saw effects of all three test compounds, and Lab 3 saw effects of 2-chlornitrobenzene and methoxychlor, but not DE-71. Some reasons for discrepancies have been given in the foregoing discussion. This does not appear to this reviewer like a high degree of replicability/transferability. In addition, the lack of reproducibility of the other parameters raises the question of why all of these other parameters were measured, and then, why the pubertal assay is being developed instead of more straightforward tests.

Overview:

I completely agree with the "strengths" section of the ISR (page 73) that an *in vivo* screen is desirable. I do not agree that this assay has as much strength as is discussed in the ISR. I do not agree with the conclusion that "The current study demonstrates that the female pubertal protocol is transferable and reproducible in contract laboratories." As indicated above, day of vaginal opening was reasonably transferable, but many of the other parameters were not. Not being a toxicologist, I do not know what level of replication from lab-to-lab is expected. From an endocrinological point of view, a well-controlled study should be entirely (or at least nearly entirely) repeatable from lab-to-lab.

Although not the task I was assigned, I will make an unsolicited statement. This protocol is very disappointing from an endocrinological point of view, and although it addresses endocrinological questions, it appears to have been developed primarily by toxicologists without sufficient input from experts in endocrinology. My opinion is that the protocol could have benefited from the inclusion of at least reproductive and developmental endocrinologists in its development. My personal assessment is that, in its current form, it will provide scant information relative to the amount of work that will go into the experiments. As I have indicated, much of the work would not be publishable in a reputable endocrine journal. While probably not my place, I recommend that a group of scientists with diverse expertise (from toxicology to reproductive and thyroid physiology and endocrinology) be convened in the style of a scientific network to discuss this

protocol from a wide range of perspectives. To do it serially, as is being done, slows down the process.

REFERENCES

1. Becker, J.B., Arnold, A.P., Berkley, K.J., Blaustein, J.D., Eckel, L.A., Hampson, E., Herman, J.P., Marts, S., Sadee, W., Steiner, M., Taylor, J., Young, E., 2005. Strategies and methods for research on sex differences in brain and behavior. Endocrinology 146**,** 1650-1673.

2. Howdeshell, K.L., Peterman, P.H., Judy, B.M., Taylor, J.A., Orazio, C.E., Ruhlen, R.L., Vom Saal, F.S., Welshons, W.V., 2003. Bisphenol A is released from used polycarbonate animal cages into water at room temperature. Environ. Health Perspect. 111**,** 1180-1187.

3. Weaver, I.C., Cervoni, N., Champagne, F.A., D'Alessio, A.C., Sharma, S., Seckl, J.R., Dymov, S., Szyf, M., Meaney, M.J., 2004. Epigenetic programming by maternal behavior. Nat. Neurosci. 7**,** 847-854.

Heather Patisaul: References

Alworth, L. C., et al., 2002. Uterine responsiveness to estradiol and DNA methylation are altered by fetal exposure to diethylstilbestrol and methoxychlor in CD-1 mice: effects of low versus high doses. Toxicol Appl Pharmacol. 183**,** 10-22.

Gallo, D., et al., 1999. Reproductive effects of dietary soy in female Wistar rats. Food Chem Toxicol. 37**,** 493-502.

Gioiosa, L., et al., 2007. Developmental exposure to low-dose estrogenic endocrine disruptors alters sex differences in exploration and emotional responses in mice. Horm Behav. 52**,** 307-16.

Goodman, J. E., et al., 2006. An updated weight of the evidence evaluation of reproductive and developmental effects of low doses of bisphenol A. Critical Reviews in Toxicology. 36**,** 387-457.

Kato, H., et al., 2003. Changes in reproductive organs of female rats treated with bisphenol A during the neonatal period. Reprod Toxicol. 17**,** 283-8.

Marcondes, F. K., et al., 2002. Determination of the estrous cycle phases of rats: some helpful considerations. Braz J Biol. 62**,** 609-14.

Patisaul, H. B., Polston, E. K., 2007. Influence of endocrine active compounds on the developing rodent brain. Brain Res Rev.

Rubin, B. S., et al., 2006. Evidence of altered brain sexual differentiation in mice exposed perinatally to low, environmentally relevant levels of bisphenol A. Endocrinology. 147**,** 3681-91.

Rubin, B. S., et al., 2001. Perinatal exposure to low doses of bisphenol A affects body weight, patterns of estrous cyclicity, and plasma LH levels. Environ Health Perspect. 109**,** 675-80.

vom Saal, F. S., 2006. Bisphenol A eliminates brain and behavior sex dimorphisms in mice: how low can you go? Endocrinology. 147**,** 3679-80.

vom Saal, F. S., Hughes, C., 2005. An extensive new literature concerning low-dose effects of bisphenol A shows the need for a new risk assessment. Environ Health Perspect. 113**,** 926-33.

Whitten, P. L., et al., 1993. A phytoestrogen diet induces the premature anovulatory syndorme in lactationally exposed female rats. Biology of Reproduction. 49**,** 1117-1121.

Yener, T., et al., 2007. Determination of oestrous cycle of the rats by direct examination: how reliable? Anat Histol Embryol. 36**,** 75-7.

3.0 PEER REVIEW COMMENTS ORGANIZED BY REVIEWER

Peer review comments received for the female 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 Jeffrey Blaustein Review Comments

Pubertal Female Rat Assay Review

1. Clarity of stated purpose of the assay.

The purpose is clear.

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

Data interpretation: It is unclear which data interpretation this refers to, since data interpretation is done at various stages… by the contract lab and by EPA. Having said that, clarity of the expected data interpretation may not be optimal. There a number of statistical issues to be considered in data interpretation. Many of the statements referred to, for example, an increase that was not significant. Statistically, an increase that is not significant is not an increase, and should not be phrased as though it is. Similarly, in the Summary Pubertal Interlab Results document, blue cells highlight "apparent," not statistical, dose-response relationships. Why? In biological research, all that counts is statistically significant effects. More statistical issues are discussed in section 4.d.

To be truly objective about the value of the work and to be statistically correct, the interpretation should rely on good statistical practices. The ISR in places has an appearance of wanting to "prove" the hypothesis/conclusion that this protocol is transferable.

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

If the purpose of the assay is to quantify the effects of chemicals on pubertal development and thyroid function, then the procedures should optimize the chance of success and minimize confounds that would obscure the results. This reviewer sees a number of serious problems with the protocol that present confounds.

Supply of animals for the experiment and confound of stress exposure: First and foremost, this reviewer has a major concern in the way that the animals are received. In Section IV, paragraph 2 of the protocol, it is stated that rats are "bred in-house or purchased from a supplier as "timed pregnant" dams with arrival at the laboratory on gestation day (GD) 7, 8, 9 or 10". The use of timed pregnant animals in a reproductive study, or for that matter in most studies, is contraindicated, because shipping is a stressor, and gestation is a time of vulnerability to stress for both the fetus and the mother. Therefore, this introduces a major confound in the protocol. Depending on how and when rats are supplied (shipped "timed-pregnant" vs. bred in lab, some animals will not be exposed to a stressor, some to a major stressor. In addition, the developmental age prenatally that the rats are exposed to the stressor will vary depending upon day of pregnancy that the rats are shipped. It is likely, but should be determined if this is a confounding factor or not, that prenatal stress influences the fetus's physiology. If there are influences, which is this reviewer's expectation, then use of "timed pregnant" females should be considered unacceptable. In-house breeding is complicated and may require additional facilities that some contract laboratories have. Therefore, it may decrease the number of laboratories equipped to do the experiments. However, use of timed pregnant animals is a serious flaw in the design. If the experiments were submitted to an endocrine journal of which I am editor, they would not be accepted.

A secondary problem with use of animals derived from mothers shipped while pregnant is the possibility that the stress of shipping compromises maternal care of the F1 generation. Since quality of maternal care is a prerequisite to normal development, and suboptimal maternal care can have epigenetic effects on the offsprings' subsequent response to hormones (Weaver et al., 2004), and perhaps xenoestrogens, use of timed-pregnant rats presents a major problem. There are so many interactions between the hypothalamo-pituitary-adrenal axis and the hypothalamo-pituitary-gonadal axis that the influence of stressors on the dependent variables has to be considered in the design of the protocol.

Necropsy. Test Method, Section X. para 1. It is stated that "On the day of kills, moving the cages or otherwise stressing the animals unnecessarily should be avoided so that variations in stress-related hormone levels are minimized." Although stress-related hormones are not being assayed in this study, this statement should still be more directive. There is no need to move or clean cages on the day of euthanasia, and it should be prohibited.

General Statement about stressors: The influence of potential sources of stressors really needs serious consideration by an expert in developmental influences of stress on physiology, and not just the influence of stress on stress-related hormones.

Diet: The ISR (page 78-79) makes the argument that, although phytoestrogens in the diet may have effects on vaginal opening, this is unlikely to be a concern, because control groups will be exposed to the phytoestrogens as well. This demonstrates a lack of understanding of physiology. Just for the sake of argument, take the case of the presence of a hypothetical antiestrogen with no estrogenic effects in the feed. The antiestrogen might be expected to block the effects of an estrogenic test compound, but be without effect in the control group. This would then lead to a false negative. One can come up with all sorts of scenarios in which a compound in the feed would influence the experimentals, but not the controls. For example, a drug that has a permissive effect on action of an estrogen. It would have no effect in controls, but a very dramatic effect in the experimentals. I must respectfully disagree with the conclusion that it is "prudent to set a limit on the concentration of phytoestrogens in feed used in the pubertal assay." The cost for requiring that phytoestrogen-free or at least quite reduced would seem to be minimal contrasted with the cost of each of these studies.

Animal housing. Animals are housed in clear plastic cages. Problems with leaching of bisphenol A from some plastic cages have been documented (Howdeshell et al., 2003). According to these authors, polycarbonate and polysulfone cages leach bisphenol A, but polypropylene cages do not. Since this presents another potential confound, the use of particular plastics and methods for cleaning them should be given a great deal of thought, and the protocol should be very proscriptive in what is acceptable.

Littermate effects. Test Method, Section IV, para 4: It is stated "Avoid placing littermates in the same group." Since litter-effects can be so robust, this statement should be considerably stronger than it is. If the experiment is worth doing, then the protocol should be very clear that placing littermates in the same group is unacceptable.

Test Method, Section IV, para 6: The second statement is very ambiguous, and should be more directive. In addition, the statement about littermates states that "littermates should not be in the same group." Since animals were assigned to groups in paragraph 4, this seems misplaced and therefore a possible source of confusion.

Injection vehicle. Test Method, Section VI, para 2. Corn oil is the preferred vehicle. However, it is not stated that this must be of pharmaceutical grade. Relying on the experimenter to make judgments of clarity, sedimentation and odor, without specifying a common source or grade will likely lead to differences among laboratories. Likewise, leaving up to the lab the choice of corn oil, water or carboxymethylcellulose is a mistake. Although this reviewer is not a pharmacokinetics expert, the solvent would be likely to influence the rate of uptake into the circulatory system. In addition, more information needs to be given regarding the method of making up of the solutions. Should the solutions be warmed or not? Should they be subjected to sonication? Etc.

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

a. comprehend the objective

The statement of purpose and applicability is quite clear.

One small exception: Test Method, Section VII, para 1. Replace phrase "*in extremis*" by « at the point of death, » since this Latin term is not understood by all. It is also in the wrong place in the sentence, since as written, it states that they will be euthanized to the point of death, and it should probably state that the animals are found in the cage near the point of death should be euthanized.

b. conduct the assay

There are a number of problems in the description of the protocol that are likely to incorrect procedure. In general, I do not think it is sufficiently proscriptive and leaves too much to the judgment of the experimenter. Why should any aspects of the protocol be left to the discretion of the experimenter. This would be likely to result in variability in results.

Water: It is stated that tap water in not acceptable, and deionized water is preferred. Since all labs have access to deionized water, why not simply require it to standardize the methods as much as possible? As with food, there is evidence that some water supplies can contain compounds with estrogenic properties. It would be most prudent to state requirements.

***Decapitation*:** Test method, Section X. para 2. It is stated that the preferred method of kill is by decapitation without any form of anesthesia. More discussion is needed, since this statement conflicts with the statement in the Guide for the Care and Use of Laboratory Animals:

"Euthanasia is the act of killing animals by methods that induce rapid unconsciousness and death without pain or distress. Unless a deviation is justified for scientific or medical reasons, methods should be consistent with the *1993 Report of the AVMA Panel on Euthanasia* (AVMA 1993 or later editions). In evaluating the appropriateness of methods, some of the criteria that should be considered are ability to induce loss of consciousness and death with no or only momentary pain, distress, or anxiety; reliability; nonreversibility; time required to induce unconsciousness; species and age limitations; compatibility with research objectives; and safety of and emotional effect on personnel.

Euthanasia might be necessary at the end of a protocol or as a means to relieve pain or distress that cannot be alleviated by analgesics, sedatives, or other treatments. Protocols should include criteria for initiating euthanasia, such as degree of a physical or behavioral deficit or tumor size, that will enable a prompt decision to be made by the veterinarian and the investigator to ensure that the end point is humane and the objective of the protocol is achieved.

Euthanasia should be carried out in a manner that avoids animal distress. In some cases, vocalization and release of pheromones occur during induction of unconsciousness. For that reason, other animals should not be present when euthanasia is performed.

The selection of specific agents and methods for euthanasia will depend on the species involved and the objectives of the protocol. Generally, inhalant or noninhalant chemical agents (such as barbiturates, nonexplosive inhalant anesthetics, and CO_2) are preferable to physical methods (such as cervical dislocation, decapitation, and use of a penetrating captive bolt). However, scientific considerations might preclude the use of chemical agents for some protocols. All methods of euthanasia should be reviewed and approved by the IACUC.

It is essential that euthanasia be performed by personnel who are skilled in methods for the species in question and that it be performed in a professional and compassionate manner. Death should be confirmed by personnel who can recognize cessation of vital signs in the species being euthanatized. Euthanatizing animals is psychologically difficult for some animal-care, veterinary, and research personnel, particularly if they are involved in performing euthanasia repetitively or if they have become emotionally attached to the animals being euthanatized.

When delegating euthanasia responsibilities, supervisors should be aware of this as a p0tential problem for some employees or students."

Therefore, while decapitation would be acceptable in this case, more discussion is needed before referring to it as "The preferred method…" In addition, as discussed in the "Guide for the Care and Use…" this is not something left to untrained personnel without regard to the animals' and the technician's welfare.

***Vaginal smears*:** There are many issues in doing vaginal smears that must be considered and are not covered in the protocol (Becker et al., 2005). First and foremost, there is no discussion of how this is to be done. The technique of vaginal lavage is a bit of a craft. If the cervix is stimulated, the animal enters pseudopregnancy (aka the progestational state), an anestrous period of twice daily surges of prolactin, rescue of the corpus luteum, and elevated progesterone levels. The females do not cycle, and the vaginal smear would look like diestrous stage. Doing these by an inexperienced technician without knowledge of the problems will result in pseudopregnancy, which would confound the results of effects of xenoestrogens.

In addition, reading the slides also that takes some practice. These are not all-or-none of one cell type or another. Typically, sample photomicrographs are included in protocols to facilitate the task of the technician and to make the assessment of cell type more repeatable (Becker et al., 2005). There are also times during the light: dark cycle that result in greatest consistency, since for example, the proestrous stage of the cycle lasts only 12-14 hours (Becker et al., 2005).

The ISR (page 75) states that "regularity of cycling should be given more weight than lack of statistical significance for the difference in weight of ovary or uterus in treated animals compared to controls." It is unclear why cyclicity data seem to have been dropped, and wee not included in the summary table 30. I could not find the data or discussion of it elsewhere in the report, except in discussion of the Therimmune multi-dose study. Why was it not included, if it is in the protocol?

The ISR indicates that listing vaginal cycles as "regular" or not offered an informed summary of the data. I could not find any indication on how "regular" cycles were determined, nor what the definition was for "cycling." Although page 9 of the protocol indicates how cycle length was to be computed, it does not indicate how many cycles are needed to qualify as having cycles, nor what constitutes regular, nor how to deal with the first days, which are usually

acyclic. These are complex issues. It was very surprising that daily treatment with a fairly potent estradiol did not lead to disruption of estrous cyclicity.
On page 557 of the Therimmune final report 7244-600, for example, on page 557, animal number 9186 shows ten straight days of a diestrous vaginal smear, yet on page 558 she is referred to as cyclic. She actually exemplifies acyclicity. I did not go through all of the data for examples like this, because it is clear that the instructions did not indicate precise definitions.

Minor issues with protocol

Dissection: Test Method, Section X, para 5. "The uterus is then place on a paper towel." Usually filter paper is used, since it does not stick to the wet uterus.

Test Method, Section X, para 5. "Measures to prevent drying out may be necessary if such organs cannot be weighed immediately." Protocol should state how drying out will be prevented, and should be very directive that any drying is unacceptable. In any laboratory that I know of, these would be weighed immediately at the time of dissection. The fact that they were allowed to dry out by at least one of the contract laboratories indicates to me that these laboratories do not always fully understand standard laboratory procedure. If they are to be used, they must be told details in very specific detail.

Problems with uterine weights, etc. Test Method, Page 17, para 4. Uterine weights and ovarian weights are not meaningful. The changes over the estrous cycle are likely to outweigh any effects of treatment. Likewise body weight has the same problem. An alternative would be to have parallel groups that are euthanized prior to puberty, so that the effects of the xenoestrogens on these variables can be assessed in the absence of ovarian hormones.

Test Method, Section X, para 10. If animal is not cycling, uterus and ovary will be either heavy or low, depending on stage (actually depending on the steroid hormone profile of the animal in the preceding day, so this is not meaningful. All that is important is age at vaginal opening, and perhaps whether the animals are cycling.

Test Method, Section XVI., para 2. It is stated that "body weight loss that does not exceed approximately 10 % is an indication that MTD was approached but not exceeded." It cannot be over-emphasized that since reduction in body weight, and in particular body fat, is an outcome of estradiol treatment, body weight loss is not useful as an indicator of MTD.

Test Method, Page 17, para 5. The term "estrus" is used as in "age at first estrus"; because this is ambiguous, and often refers to behavioral estrus, it should state "vaginal estrus."

c. observe and measure prescribed endpoints

As stated in section 4b, there are a number of problems that are not discussed that could lead to problems in measuring the endpoints prescribed in the protocol.

d. compile and prepare data for statistical analyses, and e. report results

Statistical issues. Statements like (ISR, page 23, line 8) "The ovarian weight was reduced, although this effect was present at p=.06, marginally beyond the p<0.05 cut-off for statistical significance" are inappropriate. Either a result is statistically significant (< .05), or it is not. Results that are not statistically significant cannot be used as support. Likewise, statements like "There was a nonsignificant increase in uterine weight (p = 0.08 adjusted for weaning), and a non-significant decrease in pituitary weight (P = 0.10)" (ISR, page 26, lin 16) are inappropriate. If the result is not statistically significant, and it is therefore due to chance, it really does not matter if it is an increase or a decrease; stating as the report does, that the non-significant increase is consistent with another paper's non-significant increase is inappropriate.

The ISR (page 68) states in reference to the methoxychlor data that "…all three laboratories did identify a similar pattern of response for this weak estrogen and this response was positive for interaction with the endocrine system at the same dose level." While this was true for vaginal opening, it was not true for ovaries, pituitary, liver, adrenal, but most importantly, the prototypical estrogen-dependent tissue, the uterus. As stated elsewhere, this indicates that either the protocol cannot be followed reliably or more likely that many of the parameters do not show reliable responses in animals treated in the way that they are in this protocol (e.g., offspring of mothers shipped during mid-gestation, ovaries intact so possibly cycling, tissues taken at a time that they are still potentially responding directly to the compounds).

The ISR (page 73) indicates that "for the chemical which had never been tested before, results were consistent across all three laboratories. This reviewer does not see data to support this statement. Just examining yellow highlights on table 30, one can see that, while this is reasonably true for vaginal opening and liver weight, there were mismatches in ovarian weight, pituitary weight, uterine dry weight and adrenal weight. This reviewer's interpretation is that either the end-points measured are not meaningful in tests of these compounds, or the protocol

cannot be reliably followed with sufficient sensitivity of outcome measurements to be meaningful.

Adjustment for weight at weaning. The basis for adjusting organ weights at termination of the study for covariance with body weight at weaning is something that this reviewer does not understand. Since the groups are matched for body weight at weaning, this should relatively be a constant. Furthermore, to make this adjustment, one would have to know that weight at weaning is highly predictive of body weight at 42 days of age in untreated rats.

Reporting of results. In order to facilitate an understanding of the effects, standardized bar graphs should be required with a specific format. Generally, control group with mean +/- some indicator of variance, followed by the same for each of the treatments. Standard indicators of significance as would be found in a journal article should be placed above and below the bars. Although the TherImmune report had bar graphs, they were poorly set up. RTI had only very complex tables to wade through. In addition, it is standard to convert values like 0.0087 grams to 8.7 milligrams to facilitate digestion of the numbers.

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

In this reviewer's opinion, this assay will add very little to a battery. For the cost involved, I would think that more direct tests of, for example, estrogenicity could be used. For example, *a uterotropic assay in ovariectomized or prepubertal rats or mice eliminates many of the confounds described in this review*.

6. Impacts of the choice of:

a. test substances

Positive controls compounds. A positive control with results that are known with certainty should be used. This is essential to demonstrate that the laboratory has the expertise and laboratory conditions sufficient to support replicating a previous result. Although a high dose can be used as a secondary control, a low dose, positive control to demonstrate reliability of the laboratory should be included in the protocol.

Dose of compounds. Test Method, Section V, para 1. Two dose levels are used with the maximum tolerated dose (MTD) chosen based on a reduction of no greater than 10 % of the mean body weight for controls. Since compounds with estrogenic activity are being

investigated, and since estrogens have anorectic and body weight suppressive effects, a different factor than MTD needs to be used.

In addition, these choices of doses are odd for endocrinological work. Since effects can differ at various points on the dose-response curve, usually a pilot dose-response curve is run. While the meaning of the term that is used in endocrinology… physiological levels… might not have meaning, can dosages be chosen based on reasonable expectation of level of exposure. The choice of two doses leaves a great deal to be desired, since most physiological work requires a dose-response curve.

Maximum tolerated dose: The ISR makes a number of statements such as the following (ISR, page 68): "Laboratories 2 and 3 reported that terminal body weight was significantly reduced in the high dose methoxychlor rats, both finding that this group weighed 94.1 percent of the controls. This indicates that the Maximum Tolerated Dose was reached, but that the body weight decrease compared t controls was not so severe as to interfere with endocrine endpoints." The use of body weight data like these to draw conclusions of maximum tolerated dose are not meaningful within the context of an experiment examining the effects of a potential estrogen, since treatment with estradiol itself can suppress body weight by 20 %, and is responsible for quite dramatic shifts over the four day estrous cycle. The animals, who have lost weight, are not ill, and their body weight loss has nothing to do with maximum tolerated dose. They are not losing weight because of toxicity; this is just a normal physiological response to estradiol, so would seem to be unrelated to the toxicological term, MTD.

In discussion of specificity, the ISR mentions (page 82) that "a good faith effort was made to identify a chemical that was both toxic to other systems but without endocrine effects." It is not surprising that one could not be found, because a toxic compound will decrease body weight, and this seems to be required to demonstrate that the dose has exceeded the MTD. However, body weight loss due to toxicity would likely be accompanied by a decrease in nutrient intake. From a physiological point of view, food deprivation causes reproductive dysfunction. Unfortunately, approaching a problem like this from a toxicological view-point with little regard to the underlying endocrinology/physiology has problems. In short, any compound that compromises nutrition would be expected to have endocrine effects.

b. analytical methods, and

Problems with uterine weights, etc. Test Method, P 17. Uterine weights, ovarian weights and body weights are rather meaningless in the context of this particular protocol, evidenced by many of the "acceptable ranges" given in the protocol's performance criteria. Normal physiological fluctuation over the estrous cycle in response to cyclic changes in ovarian hormones is likely to outweigh many effects of treatment. Although the protocol states that regularity of cycling should be given more weight than lack of statistical significance for the difference in weight of ovary or uterus in treated animals compared to controls, this begs the question of what a positive result is likely to be attributable to. Uterine and body weights are very informative within the context of an ovariectomized animal, but not in an ovary-intact. As pointed out in the Integrated Summary Report, "It is also important to note that the variation in the uterine weighs was expected, since uterine weights fluctuate during the estrous cycle and these females were killed on various days of their cycles."

Equally important, since animals are killed on PND 42, the last day of treatment, weights of any estrogen-responsive tissue or end-point will be a hodge-podge of direct estrogenic or antiestrogenic effects of the compound, indirect effects of the compound on the estrous cycle (Does it result in increased estradiol secretion? Does it result in long periods of anestrus?), and stage of the estrous cycle if the animals are indeed cycling. How can one know if a particular compound causes an increase in for example, uterine weight or a decrease in body weight *via* the compound, which is still available, having its uterotropic effect or body weight reducing effect, or by influencing these outcomes secondarily to perturbation of the hypothalamo-pituitary-gonadal axis (or by some other system)? There is a reason why the vast majority of endocrine studies on sex hormones is done in gonadectomized animals, and this is it. Considering the methods used, all that is important is whether the rats are cycling and age of vaginal opening, etc. But even with these variables, it must be remembered that advancing the age of vaginal opening can occur by a direct effect of an estrogenic compound on the vagina, or it can be secondary to HPG dysregulation.

An estrogenic xenoestrogen will advance puberty, as will estradiol, but it will also increase weight of uterus, and it will decrease estrous cyclicity. Therefore, an animal can develop a high uterine weight in a number of ways, for example, by an estrogenic xenoestrogens acting directly on the uterus (*e.g.*, methoxychlor) or by lack of cyclicity with the ovary stalled in stage of follicular development (high endogenous estradiol).

Similarly, vaginal opening is an estrogen-dependent process, so this is really predominantly a bioassay for estrogenicity. There are, however, better, straightforward bioassays for estrogenicity (*e.g.*, uterotropic assays).

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

For reasons discussed above, the statistical analysis has problems.

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

Repeatability is unlikely in many cases, because of the design of the experiments. There are simply too many confounds, ranging from differential exposure to stress conditions of some animals, food with potentially high levels of xenoestrogens, possible exposure to xenoestrogens from caging to tremendous variability because these are females, some of whom are cycling (and killed at random times in their very volatile cycle), some of whom are not. These have been covered extensively in previous sections.

Therimmune study issues:

It was not clear to this reviewer, if TherImmune was supposed to be following the "test method" supplied, or a completely different protocol. There were many differences, which could compromise repeatability.

In the Therimmune study, 1143-103 and 1143-101, page 12, timed pregnant rats were received on GD 12. The protocol states protocol states that the animals could be rec'd on GD 7, 8, 9, or 10. I do not know if this was a change in protocol, or lack of attention by Therimmune.

In response to my question to the EPA re: what the difference was between 1143-101 and 1143-103, we were told that the dosing period of 1143-103 was one week later than 1143-101. Since the animals were timed-pregnant and all animals were received (according to the protocols) on the same day, and the protocol called for starting the experiment on the same postnatal day, how can that be? If the dosing started one week later, then the animals would be one week older than the protocol required. If this is the case, and I have no way of determining this, it would suggest to me that the laboratory does not understand the importance of sticking to the protocol.

Although the protocol states limits of genistein allowed in the diet, Teklad 7012C was used, which is not routinely analyzed for genistein, and the experimenters did not have this analyzed. Furthermore, because phytoestrogens in the diet are a potential source of confound, the protocol could be much more directive. I see no reason that a specific diet could not be required.

Page 15: It is stated that the ethynyl estradiol was dissolved in ethanol prior to dilution in corn oil. It is essential that the amount of ethanol be stated. In fact, since it is not in the protocol, it should not have been part of the procedure, because it may have resulted in the presence of ethanol in one group, and not the others.

Figure 2. Figures of means are unacceptable without an index of variance.

The protocol calls for two rats per cage; they used 3/cage, but it is possible that the protocol changed.

Additional issues with items referred to in the Integrated Summary Report

It is stated that they showed transferability, because all of the drugs had expected results on advancing or delaying puberty. However, tamoxifen advanced, so it is unclear why this is considered success. While it was fine for day of vaginal opening, what would be considered positive results for the other parameters?

On page 23, line 20, it is stated that large variations in weights were expected because the animals were killed on different stages of the estrous cycle. This is correct, but if so, why bother going to the trouble and expense of collecting all of these weights?

On Page 27, it is indicated that the experiment showed transferability of the protocol for methoxychlor. While it is true that the transferability of the protocol for vaginal opening was shown, they failed to find significant effects on two other parameters which would have been expected (uterine and ovary weight). My conclusion from the data would be that they did *not* replicate two previous findings. On page 42, it is stated that ethynyl estradiol-treated rats had normal cycles. This is quite unexpected, since chronic estradiol treatment should stop the cycling by negative feedback. So, why would ethynyl estradiol result in normal estrous cycles?

Without going into great detail, my general sense from scanning the spreadsheet and Table 30 was that, although there was reasonable consistency among labs in day of vaginal opening, there was not a great deal of consistency among labs in many of the other parameters. But even in “age at vaginal opening,” Lab 1 saw no effect of DE-71 or 2-chloronitrobenzene, but saw an

effect of methoxychlor, Lab 2 saw effects of all three test compounds, and Lab 3 saw effects of 2-chlornitrobenzene and methoxychlor, but not DE-71. Some reasons for discrepancies have been given in the foregoing discussion. This does not appear to this reviewer like a high degree of replicability/transferability. In addition, the lack of reproducibility of the other parameters raises the question of why all of these other parameters were measured, and then, why the pubertal assay is being developed instead of more straightforward tests.

Overview:

I completely agree with the "strengths" section of the ISR (page 73) that an *in vivo* screen is desirable. I do not agree that this assay has as much strength as is discussed in the ISR. I do not agree with the conclusion that "The current study demonstrates that the female pubertal protocol is transferable and reproducible in contract laboratories." As indicated above, day of vaginal opening was reasonably transferable, but many of the other parameters were not. Not being a toxicologist, I do not know what level of replication from lab-to-lab is expected. From an endocrinological point of view, a well-controlled study should be entirely (or at least nearly entirely) repeatable from lab-to-lab.

Although not the task I was assigned, I will make an unsolicited statement. This protocol is very disappointing from an endocrinological point of view, and although it addresses endocrinological questions, it appears to have been developed primarily by toxicologists without sufficient input from experts in endocrinology. My opinion is that the protocol could have benefited from the inclusion of at least reproductive and developmental endocrinologists in its development. My personal assessment is that, in its current form, it will provide scant information relative to the amount of work that will go into the experiments. As I have indicated, much of the work would not be publishable in a reputable endocrine journal. While probably not my place, I recommend that a group of scientists with diverse expertise (from toxicology to reproductive and thyroid physiology and endocrinology) be convened in the style of a scientific network to discuss this protocol from a wide range of perspectives. To do it serially, as is being done, slows down the process.

REFERENCES

1. Becker, J.B., Arnold, A.P., Berkley, K.J., Blaustein, J.D., Eckel, L.A., Hampson, E., Herman, J.P., Marts, S., Sadee, W., Steiner, M., Taylor, J., Young, E., 2005. Strategies and methods for research on sex differences in brain and behavior. Endocrinology 146**,** 1650-1673.

2. Howdeshell, K.L., Peterman, P.H., Judy, B.M., Taylor, J.A., Orazio, C.E., Ruhlen, R.L., Vom Saal, F.S., Welshons, W.V., 2003. Bisphenol A is released from used polycarbonate animal cages into water at room temperature. Environ. Health Perspect. 111**,** 1180-1187.

3. Weaver, I.C., Cervoni, N., Champagne, F.A., D'Alessio, A.C., Sharma, S., Seckl, J.R., Dymov, S., Szyf, M., Meaney, M.J., 2004. Epigenetic programming by maternal behavior. Nat. Neurosci. 7**,** 847-854.

3.2 Barry Delclos Review Comments

1. Clarity of stated purpose of the assay.

The statement of purpose of the assay is reasonably clear. A revision to the initial sentence to state more specifically what will actually be accomplished with this assay might be appropriate. For example: "The purpose of this protocol is to identify chemicals that affect, after oral administration, pubertal development and thyroid function in the intact juvenile/peripubertal female rat."

One thing that is unclear is why the oral route was selected as the only possible route of exposure for this assay. While there is likely to be limited information on the majority of chemicals that will be tested in the Endocrine Disruptor Screening Program (EDSP), in cases where either planned use or pharmacokinetic data are available, it would seem that, at least in some cases, other routes might be preferred. If the goal of the assay is primarily to detect endocrine system activity that would then be further investigated and defined in higher tier studies, it would seem that a route that results in higher systemic exposure to the test chemical might be selected even if that route is not the major route of exposure to be expected in humans.

At any rate, a statement of why the oral route is specified (e.g., because it is likely to be the primary route of human exposure, etc.) would be appropriate.

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

In general, the interpretations of the data generated in the validation studies of the female pubertal protocol that are presented in the Integrated Summary Report (ISR) were clear and thorough. There were a few points that caused some confusion.

1) In the ISR discussion of the results of the validation study that included bisphenol A (ISR, page 28, lines 21-25) it is basically concluded that the assay did not detect BPA activity with the possible exception of the body weight depression. In later discussion of the effects of estrogens on body weight (ISR, page 53, paragraph starting on line 22) it appears that, based on results with ethynyl estradiol and methoxychlor, a body weight depression in the absence of an effect on other endpoints such as vaginal opening would not be interpreted as an estrogenic activity. It is not clear how the BPA data would have been interpreted for a compound that did not have the extensive body of published data that exists for BPA. If possible, a more definitive statement should be made about the sensitivity of body weight relative to the other endpoints with regard to estrogenic test compounds.

2) With regard to the discussion of BPA in the ISR, it is indicated on lines 18-19 of page 28 that body weight at vaginal opening was increased at the high dose. However, the data in Table 12 and the data presented in the laboratory study report seem to indicate that the body weight at vaginal opening was decreased at the high dose. Consideration of this result along with the ovarian weight data would likely change the conclusion reached on line 19 of page 29 that the expected estrogenicity of BPA was not detected.

3) Another instance where the ISR was in disagreement with the laboratory study report was the case of 2-chloronitrobenzene (2-CNB). The Argus report (Appendix 13) indicated that the thyroid showed histological changes consistent with a hypothyroid state, but that there were no significant changes in thyroid hormone levels. The ISR (Table 30, page 71) indicates that there was a significant decrease in T4 in the Argus study. The ISR does not mention histopathological results from the 2-CNB study thyroid component and there seems to be only sporadic use of histopathology results in interpreting the various validation studies throughout the ISR. In the case of 2-CNB, it is interesting to note that the laboratory that did

not report changes in thyroid hormone changes did report treatment-related histological changes while the other laboratories reported thyroid hormone changes in the absence of histological changes. None of the laboratories reported effects on thyroid weight for this compound. One of the laboratory study reports (Appendix 15) indicated that the meaning of an isolated rise in TSH in the absence of other significant thyroid effects was unclear. The variability of hormone measurements in the thyroid assay was discussed in the ISR, but further discussion of what will constitute a positive call when results from only one laboratory are available might be helpful.

4) In the protocol (page 17), it is indicated that changes in pituitary, liver, and kidney weights should be interpreted as relevant only if there is a significant change in organ weight relative to body weights. From the data summaries and interpretation in the ISR, it was difficult to determine if this requirement was met in all cases where changes in these particular organ weights were discussed.

5) There are several instances in the ISR (for example, pages 26, 30, 31, 50, and 70) where changes in means that are not statistically significant are discussed when the changes are consistent with what was expected based on previous studies or are consistent with other observations in the study. Comment on how such results would be used in reaching a decision on the activity of a compound might be appropriate in Section XVI (Data Interpretation) of the protocol.

6) One of the required endpoints, age at first estrus, is discussed for some of the validation studies, but not for all. For example, the discussion of the interlaboratory study does not mention how this endpoint was affected or how the result impacted the conclusions for the test compounds.

7) The detection of estrogenic activity of methoxychlor at 12.5 mg/kg in the multi-dose study (Appendix 8) is cited as an example of the sensitivity of the assay (ISR page 47, lines 14-17), but the three laboratories in the interlaboratory comparison study did not detect activity at this dose. It is of interest to note that the study that did detect activity at 12.5 mg/kg appeared to use the diet with the highest phytoestrogen level, although it is certainly not clear what factors might have contributed to the discrepant result.

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

EPA has provided ample background and discussion on the biological and toxicological relevance of the assay and its ability to multiple mechanisms of interference with estrogen and thyroid hormone activities. The endpoints specified are amenable to a large scale screening program. Concern over a lack of a clear demonstration of assay specificity to this point is an issue that has been recognized and discussed. The controversial issues of sensitivity differences of rat strains and the potential impact of diet were discussed and EPA has provided reasoned explanations for their decision to recommend the Sprague-Dawley rat and to set an approximate limit of 300 ppm genistein-equivalents of phytoestrogens. These decisions will no doubt be reviewed as additional data become available. One issue that was not directly addressed in the Integrated Summary Report or the protocol itself was the question of whether confining testing to the Maximum Tolerated Dose and one half of the Maximum Tolerated Dose could miss important endocrine activity that would be evident at low doses. A footnote in the protocol to restate the EPA position on this issue should be considered.

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

a. comprehend the objective

See comment under "1. Clarity of the stated purpose of the assay" above.

b. conduct the assay

For the most part, the methodology is clearly described. I would propose the following for consideration. All pages indicated refer to the protocol.

1) In the list of endpoints under organ weights on page 1, consider indicating that the thyroid is weighed after fixation. While this is a standard practice for protocols evaluating the thyroid, highlighting this in the list of endpoints would be helpful. Also, consider indicating in Section II of the protocol that kidney histology is optional and add clinical (serum) chemistry, blood urea nitrogen and creatinine as optional endpoints.

2) Page 2, footnote 3: What is the basis for the statement that totally synthetic diets are not appropriate? Is this because data indicate that they are not suitable, or because there are insufficient data to support their use? Most studies (for example, reference 58 in the ISR) appear to use the older AIN-76

formulations, and there appear to be less data for the revised AIN-93 diet formulation. AIN-93 has been used in reproductive studies (*e.g.* Collins *et al*., Effects of flaxseed and defatted flaxseed meal on reproduction and development in rats, Food and Chemical Toxicology **41**: 819-834, 2003). In that published study, data presented for sexual development for the female controls on the AIN-93 diet seems to fall in line with performance criteria given for the female pubertal protocol.

3) Page 2, footnote 3: The recommendation of the limit on the level of phytoestrogens is based on data from the uterotrophic assay. Recent data reported by Thigpen *et al.* http://www.ehponline.org/docs/2007/10165/abstract.html indicates that the CD Sprague-Dawley rat may be insensitive to phytoestrogens relative to the F344 rat. Response to an estrogen challenge was not reported in that study, but it would appear that there would be concern with the level of phytoestrogens specified in the current protocol if strains other than the Sprague-Dawley are used in the future.

4) Page 2, footnote 5: Is the statement that the AAALAC guideline extremes are "only marginally tolerable for the pubertal assay supported by published data that could be cited?

5) Page 3: With regard to guidance on the culling of litters, it might be useful to directly state that litters smaller than 8 can be used without adjustment.

6) Page 4, section 6: Some guidance on acceptable purity might be provided here. The methoxychlor (MXC) used in some of the validation studies had a purity of less than 90%. Earlier literature indicated that technical grade MXC was more potent than highly purified MXC due to presence of active metabolites in the technical grade material. For unknown compounds going through the screen, will use of technical grade material be acceptable or is highly purified compound preferred? In any case, some statement might be provided.

7) Section IV: The diet that the timed pregnant females were fed by the supplier prior to shipment to the test laboratory should be reported.

8) Page 4, second paragraph: Guidance is given in this paragraph for setting the low dose to be used in the assay. For the interlaboratory validation study with 2-CDNB, it was not explained why the low dose selected did not follow this guidance.

9) Page 6, section X, Necropsy: The statements here in footnote 8 and on page 7 regarding carbon dioxide asphyxiation as inhumane seem to be at odds with current AVMA Panel recommendations http://www.avma.org/issues/animal_welfare/euthanasia.pdf . Those recommendations indicate that carbon dioxide is an acceptable method of euthanasia while decapitation is conditionally acceptable with scientific justification. Decapitation can be scientifically justified here, and the comments on carbon dioxide asphyxiation seem unnecessary.

c. observe and measure prescribed endpoints

1) Page 8, Section XII Histology: Add "(optional)" after "kidney".

2) Page 8, third paragraph in Section XII: There is mention of reporting changes in numbers of primary and atretic follicles. Is the ovarian evaluation intended to be a qualitative evaluation of a single section, or are step sections necessary?

d. compile and prepare data for statistical analyses

1) Page 3, last sentence of the second paragraph: It is indicated that placement of litter mates in the same group should be avoided, but the last sentence in the fifth paragraph suggests that there may be situations where this could occur. It seems that either a stronger directive should be given (i.e., Do not place litter mates in the same group) or guidance on how litter mates in the same group should be reported and handled in the statistical analysis provided.

e. report results

1) Page 5, Section VIII, Vaginal Opening: It is indicated that documentation of vaginal threads or the appearance of pin holes are important and that it is

critical that the "initiation" of vaginal opening be recorded. However, guidance on the interpretation and reporting of these endpoints is not provided.

2) Page 12, second paragraph: This is somewhat confusing given the guidance on interpreting pituitary, liver, and kidney weights given on page 17 of the protocol. There it is indicated that organ weight changes for these organs should be considered only if they change significantly relative to terminal body weight. As mentioned previously, it wasn't clear in the ISR or lab study report whether the ratios were being taken into account in the discussion of changes in the weights of these organs. It seems that for these three organs, the ratios to terminal body weight should also be reported in the table.

3) Page 17, last paragraph: I don't believe that this paragraph is necessary. Presumably, the fact that an isolated endpoint showed significance at the low dose, but not at the high dose would be taken into account in reaching a conclusion about the substance based on the results of this assay and when considered with results obtained in the other battery assays.

4) Page 18, Table 6: Consider adding a column describing anticipated changes for an estrogen antagonist.

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

EPA has provided a good discussion of the strengths and limitations of this assay. Strengths include the coverage of multiple mechanisms of action that are not covered in other proposed assays for the battery and exposure during a sensitive time period, although it could be argued that the perinatal period would be more sensitive for some endpoints. A major current limitation, as pointed out by the EPA, is the lack of demonstrated specificity of the assay. Another limitation is the impracticality of conducting the screen in more than a single strain to provide confidence in studies where endocrine activity is not detected.

6. Impacts of the choice of the following in terms of demonstrating the performance of the assay:

a. test substances

The selected test substances covered a variety of mechanisms of action purported to be detected by this assay and, for the most part, expected results were obtained. The selection of fenarimol, which apparently has a mixed mode of action that was not known at the time of selection, was described in the ISR as unfortunate, but it is likely that some unknown test compounds will interact with multiple targets and the data obtained with that compound did have utility. Inclusion of a broader range of compounds in the interlaboratory study would have provided added confidence in the performance of the assay, although it is recognized that the additional cost may have made that impractical.

b. analytical methods

1) The issue of the variability of tissue weights of the pituitary and adrenal glands has been addressed by indicating that steps need to be taken to avoid drying out prior to weighing (page 7). Weighing the organs after fixation (tissues placed in fixative immediately after removal) might also be considered as an approach to remedy the variability in these organ weights.

2) The issue of variability of hormone measurements was discussed in the ISR, and it was noted that performance criteria for TSH had not yet been established. The mean TSH levels in control animals vary widely across the various studies discussed in the ISR. The multichemical study used carbon dioxide as a method of kill (Appendix 5, page 13) and had a high control TSH level, but similar levels were seen in one of the laboratories in the interlaboratory studies where brief carbon dioxide followed by decapitation was the method of kill (Appendix 15). The two other laboratories in the interlaboratory study used the same method of kill as the latter study and both of these studies reported control levels of TSH considerably lower than those reported in Appendices 5 and 15. The Office of Research and Development (ORD) studies reported in Table 22 (page 51) of the ISR had the lowest control levels of TSH, which no doubt reflects the extensive

experience of this laboratory with this assay. Perhaps strict adherence to the ORD protocol should be specified. However, while the variability of this endpoint is problematic, it appears that high control levels did not necessarily interfere with the ability of a laboratory to detect treatment effects.

c. statistical methods

1) Page 9, to paragraph: "When statistically significant effects are observed ($p < 0.05$), treatment means are examined further using appropriate pairwise comparison tests to compare the control with each dose group." This approach may be overly conservative for a screening assay for certain multiple comparison procedures. In many of the validation study procedures, Dunnett's test was used for comparisons of treatment groups to controls. The following statement is taken from Haseman *et al*., Statistical issues in the analysis of low-dose endocrine disruptor data" Toxicol. Sci. 61: 201-210, 2001: "For example, Dunnett's test is a standalone test that does not require statistical significance of an overall ANOVA to be valid. However, many investigators who used Dunnett's test required statistical significance of an overall ANOVA before making pairwise comparisons. Because the critical values for Dunnett's test were derived without consideration of an overall ANOVA, requiring this additional significance may result in a somewhat conservative test. Specifically, there were a few instances in which our reanalysis found significant pairwise differences by Dunnett's test that were not reported as significant by the study investigators who themselves also used Dunnett's test. Such differences were apparently due to the extra requirement of a significant overall ANOVA imposed on Dunnett's test by the study investigator."

Revision of this recommendation should be considered, depending on the multiple comparisons test to be applied. While this statistical guidance may not have affected the interpretation of the validation studies, it could potentially affect conclusions in future screens.

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

In general, the results reported in the validation studies, particularly the interlaboratory study conducted with three chemicals, indicate that the protocol generates data that lead to similar ultimate conclusions. Some issues that need to be clarified in interpreting the data have been raised above under responses to the earlier questions.

3.3 David Furlow Review Comments

Peer review of EPA Female Pubertal Rat Assay

A. Review summary:

The female prepubertal rat assay seeks to develop a reproducible and sensitive screening assay for the influence of xenobiotics on endocrine related endpoints, within an intact animal. The assay is sensitive and reproducible for the panel of compounds tested, and the endopoints measured are appropriate as a Tier 1 screen in conjunction with a battery of other in vitro and in vivo tests. The major drawbacks include continued concern about rat strains chosen for the study, the window of exposure to the compounds, and the lack of behavior endpoints in the assay. Most of the detailed comments are in response to questions 4 and 5.

B. Direct answers to charge questions:

1. **Clarity of the stated purpose of the assay**. The assay's stated purpose is to test the effect of xenobiotics on the endocrine system of the prepubertal female rat, specifically with regard to estrogenic/antiestrogenic and antithyroid activity or disrupted hypothalamic and pituitary function as relates to the onset of puberty. This appears straightforward enough; however, in the peer review charges, the effect of mixtures is mentioned as an application of the assay. No instruction of how best to perform mixture analysis is provided (an increasingly important issue in toxicology), nor is it addressed in the ISR.

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

   These issues all appear appropriate; some caveats are discussed in other sections below. One item to note here is that changes in liver enzyme profile are an indicator of the presence of a potential thyrotoxicant (p. 18 Table 6), but liver enzymes are not measured in the protocol.

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

   Measurements of are deemed appropriate for measuring interference with estrogen and thyroid hormone endocrine systems. Biologically relevant endpoints for thyroid hormone such as T4 and TSH measurements and thyroid histology are appropriate, as are time to vaginal opening, parameters of the onset and length of the estrous cycle, and uterine histology for alterations in the function of the hypothalamic-pituitary- gonadal axis.

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

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

   In general, the protocol is well written, easy to follow, and instructions on data collection and reporting are clear.

   Specific comments on the test method:

   - Specific plastic used for caging (p.2 line 2) should be described in more detail. I found at least one article describing the leaching of potential endocrine disrupting chemicals from polycarbonate tubs used in rat or mouse housing, and this can be drastically increased after exposure to alkaline washing conditions and/or high temperature (Koehler et al. Lab Animal 32: 24-27, 2003; Everitt and Foster ILAR 45: 417-424 2004).

- Specific water bottles to be used for ad lib drinking are not described (see previous concerns re: housing and washing and care of plastics) (p.2 line 8).
- Within an experiment, juvenile rats should be obtained from either pregnant females from an in-house breeding or purchased from a supplier but not mixed in a study or when comparing between studies (p. 2 line 19). Inadvertent prior exposure to chemicals during gestation and the perinatal period, or maternal stressors, can influence later responses to these same or different chemicals to the offspring (e.g. Newbold et al. Reproductive Toxicology 23:290–296 (2007)). Therefore, the source of animals used in the studies ought to be standardized.
- Designation of the Maximum Tolerable Dose level is unclear to me (p. 4 line 2), since the protocol is presumably to be applied generally to compounds with both known and unknown general toxicity profiles. The way this is worded implies that preliminary studies are done to determine the MTD prior to a full scale assay for endocrine system effect, which may not actually be the case.
- One significant source of stress that can be avoided is to sacrifice the animals at an area of the laboratory away from the rest of the animals to be sacrificed that day. (p. 6 line 23).
- The assay protocol includes uterine wet weight as an endpoint (p. 7 line 7), yet in the ISR p. 59 line 8 it states that this is deleted from the protocol due to variability. This is unfortunate, since in ovariectomized rats, uterine wet weight is an excellent predictor of estrogenicity of a compound although the EPA peer review web site lists a rat uterotrophic assay as one of the battery of tests to be evaluated separately.
- The lack of a standard "recommended" assay for hormone measurements (p. 7 line 28), (several options are deemed acceptable as long as quality control samples are run) may have to be revisited due to the outlying or inconsistent T4 and TSH values reported in the interlaboratory exercise (Argus laboratory values; Figure 2 page 67 and Table 30 page 73 of the ISR). This may include having a separate lab coordinate all the hormone assays, or settling on a recommended assay kit and vendor to improve reproducibility. Nevertheless, the T4 and TSH values at least changed in the same direction in all three laboratories.

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

Strengths of the assay: The major strength of the assay is that it is able to detect the activity of chemicals with regard to the endocrine system in an *intact* mammalian system. Cell culture and *in vitro* biochemical screening assays are useful to provide information on what effect a chemical *can* have on the function of the endocrine system but does not account for

cell specific uptake, metabolism, effect of other circulating growth factors/hormones etc. On the other hand, *in vitro* studies provide more mechanistic information than are inferred from the *in vivo* system. Therefore as a component of a battery of assays, the prepubertal female rat assay appears relatively sensitive (see below for caveat) and standardizable to detect interactions between xenobiotic chemicals and estrogen and thyroid hormone physiology. It is especially noteworthy that a so-called "weak" estrogen, methoxychlor, and a "weak" thyroid disrupting compound DE-17 produced quantifiable and generally reproducible effects on the expected endpoints.

Limitations of the assay:

a. *Rat strain choice for these assays is still very much open for debate*. While it is apparent that the EPA is aware of potential strain influences in sensitivity to endocrine disrupting chemicals, the assay only proposes to use Sprague Dawley rat (Crl:CD(SD)). This choice is based largely on comparing only Wistar and Sprague-Dawley rats, and experience with the male prepubertal assay. However, there are several publications demonstrating increased sensitivity of Fisher F344 rats to estrogens and BPA relative to at least SD rats in multiple endpoints as pointed out in Appendix 12, several publications in the open literature, and as noted by the ISR. This point should not be dismissed in considering the utility of the assay based only on practical considerations.

b. *Exposure duration is confined to PND 22-PND 42*. The effect of fetal or perinatal exposure to chemicals is not accounted for in the current protocol. There is growing evidence that early exposure to chemicals or maternal stress can "program" adult physiology of the offspring (e.g. Newbold et al. Reproductive Toxicology 23:290–296 (2007)), in addition to sexually dimorphic behaviors (see c).

c. *The assay does not account for behavioral effects*. In many cases, low doses of estrogenic compounds can affect sexually dimorphic brain organization, at lower or similar doses to those required to elicit morphological changes in reproductive tissues. Therefore, I was somewhat surprised that behavioral endpoints were not included in the assay, such as lordosis, for example.

d. *The specificity of the assay remains a question due to the effect of 2-CNB.* A potential negative control, 2-CNB, delayed vaginal opening and increased TSH levels. This issue is discussed in more detail below, but may be the result of increased metabolism of estradiol

and/or thyroid hormones by the 2-CNB exposed liver. It is also interesting to note that even differences in caging can lead to significant differences in age of vaginal opening and estrous cycle parameters (Firlit and Schwartz, Biol. Reprod. 16:441-444, 1977) indicating the effect of non-specific stressors on these parameters although the exact mechanism ins not understood, to my knowledge.

6. **Impacts of the choice of test substances and methods chosen to demonstrate the performance of the assay.**
   a. test substances
   b. analytical methods
   c. statistical methods in terms of demonstrating the performance of the assay

The choice of test substances for the testing the performance of the assay are appropriate and span a spectrum of agents with distinct suspected modes of action. A few comments are included for future consideration however. For a test substance to detect thyroid hormone disruption, a more relevant choice than PTU may have been ammonium perchlorate since this compound is a known inhibitor of iodine uptake by the thyroid and is found in ground water, particularly near air force bases. Also, a more specific aromatase inhibitor such as fadrazole rather a general steroidogenesis inhibitor such as ketoconazole than may have been useful for assessing the specificity of the female rat assay in particular. Interestingly, it is apparently difficult to find a test substance known to be generally toxic but does not in some way impact any of the endocrine related endpoints in this assay. One possibility is that high doses of generally toxic chemicals induce hepatic phase I and II metabolizing enzymes and phase III transporters as a by-product of exposure. This possibility was noted in the ISR (p 70 line 19)In this regard, screening the activity and/or mRNA expression of a panel of known steroid and thyroid hormone metabolizing enzymes in liver biopsies should be an important endpoint to consider.

The statistical analyses proposed appear appropriate, but this is not my area of expertise.

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

In general, repeatability and reproducibility was satisfactory, with the exception of the T4 measurements discussed under charge question 4. Furthermore, it is not unexpected that the wet weights of tissues would have so much variance and thus be of limited value since they cannot obviously be compared to wet weights prior to treatment.

3.4 Heather Patisaul Review Comments

Critique Summary

The need for an effective pubertal assay is unequivocal. There is sufficient evidence to suggest that the proposed protocol will be able to detect a large cohort of compounds that are capable of disrupting the endocrine system and thus alter the timing of puberty. Disruption of thyroid activity will also be detected in this assay although the inclusion of these endpoints complicates the design of the assay and introduces a number of caveats as discussed in detail below. Although the assay is generally well constructed, there are a number of critical issues that diminish the biological and toxicological relevance of the assay.

The assay would benefit from simplification. A few salient endpoints, collected well and with careful controls, would be far superior to the broad spread of largely unrelated endpoints currently proposed. Inclusion of both estrogen and thyroid related endpoints complicate the interpretation of the data and as such, in many cases, the results of the data are listed as "difficult to interpret." For the assay to be effective and reliable a data set that is "difficult to interpret" should be rare rather than the norm. A screening assay should yield straightforward results that clearly and quickly identify compounds that require further testing.

A number of elements within the protocol diminish the functional utility of the assay. Ovarian and uterine weights are generally uninformative and complicated by cycle. Inclusion of the thyroid endpoints precludes the use of a needed positive control group for the pubertal measures. The duration of estrus monitoring is too short and the use of daily lavage will likely induce pseudopregnancy in some animals, potentially confounding the data. Failure to eliminate phytoestrogens introduces an unnecessary confound and increases the risk of inter-laboratory

variability. Finally, only two doses are to be used, both of which are based on body weight and neither of which will approximate a "typical" human or wildlife exposure. The failure to include a dose within a reasonably physiological range is a considerable concern.

1. Clarity of the Stated Purpose of the Assay

The purpose of the assay is clearly stated and well justified. The protocol is intended to detect alterations in sexual maturation and thyroid function by exogenous chemical exposures during the prepubertal period.

It should be noted that the majority of endpoints within the assay are related to puberty rather than thyroid function. The thyroid endpoints, although interesting in their own right, feel out of place in the context of the assay. They also complicate the interpretation of the results (as discussed more in depth below) and preclude the inclusion of a positive control for the estrogenic endpoints. A "female pubertal assay" should be specific, simple, and straightforward. If possible, it would be advisable to develop a separate assay to assess thyroid function.

2. Clarity, Comprehensiveness and Consistency of the Data Interpretation

In many cases, the results of the data are listed as "difficult to interpret." This is a significant concern as an EDSP Tier 1 Screening Assay should generate a data set that is relatively simple to interpret and reliably identifies compounds that require further testing. A laboratory running this assay should easily be able to conclude that a compound either produces or fails to produce an effect. The interpretation of the results should be as unequivocal as possible. The results of the study using 2-chloronitrobenzene (beginning on page 70 of the Integrated Summary Report) best illustrates how results from a compound, for which little about potential endocrine activity is known, might be interpreted. Two of the three laboratories reported significantly delayed vaginal opening. Significant changes in weight at vaginal opening, liver weight, adrenal weight and uterine weight were also observed in at least two of the laboratories. This was an unexpected finding but the authors ultimately conclude (albeit tenuously) that 2-chloronitrobenzene interacts with the endocrine system though in indeterminate mechanism. No information is given as to how this compound would then be classified. Would it move to Tier 2 screening? Would the results be questioned? How would this data be received by the EPA? This compound was selected for screening because it was hypothesized to have no effect on the

endocrine system. The data do not support the hypothesis. How would that data be used? Would it be questioned?

3. Biological and Toxicological Relevance

The need for an effective pubertal assay is clear and well justified. However, there are a number of design issues that diminish the efficacy and relevance of the current protocol.

a. Lack of a Positive Control Group

The lack of a positive control is a serious concern. Within the Integrated Summary Report, this omission is justified by the argument that it is highly unlikely that a single compound that will generate a positive result for all endpoints in the assay. This problem results from the inclusion of experimental endpoints designed to address two different and largely unrelated questions. By lumping pubertal endpoints, which assess estrogen action, together with thyroid endpoints, the choice of an appropriate positive control becomes complicated. It is readily apparent that the most salient and critical goal of this assay is to identify compounds that affect puberty. As such, a positive control that reliably and consistently advances puberty should be included, regardless of whether or not any thyroid endpoints are altered. Estradiol, DES, or estradiol benzoate would all be appropriate positive controls and at least one should be used by all labs for this purpose. Any labs not observing an effect with the positive control would then immediately know that they have a problem executing the assay properly.

c. Selection of Dose

Only two doses will be used in the assay. Both are based on body weight with the second being half of the first. By basing the doses used for the assay on the maximum tolerated dose (MTD), both are likely to be quite high compared to what humans and wildlife could reasonable expect to be exposed to. Some of the most concerning findings with endocrine disrupting compounds have occurred at doses that are well within the realm of human exposure and far lower than would be chosen by the parameters of this assay (Alworth et al., 2002; Gioiosa et al., 2007; Goodman et al., 2006; Kato et al., 2003; Rubin et al., 2006; Rubin et al., 2001; vom Saal, 2006; vom Saal and Hughes, 2005). One of the recommendations in the Final Report of the Endocrine Disruptors Low-Dose Peer Review (2001, NEIHS) was to replicate and validate “low dose”

studies. Although the argument for "low dose effects" is controversial, employment of a low dose in the assay is well justified and would address this issue. The data within the Integrated Summary Report illustrate the critical need for the employment of a lower dose. Atrazine, Bisphenol-A and methoxychlor all produced significant effects at substantially lower doses than what would likely be used under the current protocol guidelines. Inclusion of a "low dose" would also increase the robustness of the assay. For example, within the Integrated Summary Report the observed effects of methoxychlor at 25 and 50 mg/kg/day are argued to confirm "the transferability of the assay and provide evidence that the assay is sensitive." (Page 27, line 7-8). The use of a "low dose" as defined by the Endocrine Disruptors Low-Dose Peer Review (2001, NEIHS) or a similar protocol would significantly strengthen the assay and help to clarify whether or not "low dose" effects are of genuine concern. If left out, this controversy will continue to linger and doubt about the safety of the test compounds will remain even after they are subjected to testing with this assay.

It should be noted that most of the data regarding low dose effects have come from animals exposed during the gestational or neonatal period. Therefore it is unclear if low dose effects would be observed when the exposure begins just prior to puberty, as proposed in the current protocol. However, there is sufficient data to warrant the inclusion of a low dose.

Finally, the use of high doses may explain why, to date, no compound has produced a negative result in this assay. The highest dose to be used is defined as a statistically significant reduction in body weight with "no clinical signs of toxicity." The acceptable "signs of toxicity" are not identified or discussed in the protocol but should be, perhaps in an appendix. In general, the use of body weight to define dose is problematic for several reasons, most of which have already been addressed previously by Goldman et al, but again highlights the need for a positive control group within this assay. It is well established that estradiol administration significantly reduces body weight. Because a decrease in body weight of 10% or more can result in the disinclusion of subjects or treatment groups, the employment of a positive control group would help clarify whether or not the MTD was reached or exceeded, and whether or not the laboratory was conducing the assay properly.

4. Clarity and Conciseness of the Protocol

If the assay is to serve as a reliable and predictive tool for the identification of endocrine-disrupting compounds in females, it should be straightforward, clearly written, and easily performed in laboratories with reasonable testing experience. It should also be easily transferable and reproducible. There are a number of issues that impair these goals in the current protocol.

a. Objective – The objective is clearly stated in the protocol.

b. Methods – There are a few places within the methods where clarification is needed.

1. DIET: The diet should be free of phytoestrogens. As written, diets containing up to 300 µg/kg are allowed. The presence of phytoestrogens, even in such small quantities, needlessly impairs the sensitivity of the assay and introduces inter-laboratory variability. A number of phytoestrogen-free diets are now readily available from all of the major lab diet manufacturers so there is no reason not to exclude this potentially problematic source of endocrine disrupting compounds.

2. TRANSPORT: Animals can either be shipped timed pregnant or born in-house. This introduces gestational stress as a source of variability. Transport stress during pregnancy can delay parturition by 24-48 hours, reduce maternal care and expose the developing pups to stress hormones during gestation. All of this can affect the growth and overall well-being of the offspring. Maternal stress should be minimized as much as possible. As such, all females should be impregnated in-house with care to avoid any unnecessary disturbance. Cage changes and all other intrusions should be minimized as much as possible.

3. DEFINING THE DAY OF BIRTH: The day of birth is defined as the "morning" of PND 0. This definition is too broad. Under this definition, an animal born at 12:10 pm on a Monday could be listed as PND 0 on Tuesday and cross fostered with animals born at 11:45 am on Tuesday. This means litters could contain animals born 18-23 hours apart. This is a problem in a protocol where many of the data interpretations are to be based on body weight. A specific

time frame for the definition of PND 0, and some discussion is needed as to how animals born close to that window will be dealt with.

5. DEFINING ESTRUS: Daily lavage frequently induces pseudopregnancy in rats and can skew data regarding regularity of the estrus cycle and cycle length (Marcondes et al., 2002; Yener et al., 2007). Pseudopregnancy lasts 1-2 weeks, during which time the vaginal smear will have the appearance of persistent diestrus. No discussion as to how to deal with this frequent phenomenon is given in the assay.

6. LENGTH OF ESTRUS OBSERVATION: Animals are to be killed on PND 42. This means that estrus cyclicity will be assessed for less than two weeks and animals will be killed before they have completed the transition to full adulthood. This is an insufficient amount of time to evaluate the regularity of the estrus cycle and will likely fail to detect a number of compounds that ultimately impact endocrine action in females. The impact of endocrine disrupting compounds on the estrus cycle can be delayed by several weeks (Gallo et al., 1999; Patisaul and Polston, 2007; Rubin et al., 2001; Whitten et al., 1993). Observation of the estrus cycle for at least six weeks post puberty is strongly recommended.

7. EUTHANASIA: The use of decapitation without anesthesia is inappropriate and unnecessary. The Office of Laboratory and Animal Welfare (OLAW) within the National Institutes of Health (NIH) does not generally support the use of decapitation without anesthesia unless there are extenuating circumstances that require it. Use of CO2 asphyxiation will not alter any of the outcomes in the protocol and should therefore be used.

8. BLOOD COLLECTION: It is unclear how the blood is to be collected and prepared. The assay states that trunk blood should be collected by “inversion over a funnel” but the type of collection tube is not given. The choice of assay to measure plasma hormone levels will influence the type of collection tube to be used. In some cases, a siliconized or EDTA lined tube is required. This should be specified in the protocol.

9. HISTOLOGY: A citation for the grading scale associated with the thyroid sections is given, but no citations are listed for the evaluation of follicular development or uterine histology.

Collecting uterine and ovarian weight has consistently proven to be problematic and it is not clear how these endpoints are informative. Time of vaginal opening, time of first estrus, and regularity of the estrus cycle 10-12 weeks post puberty are far more valid and reliable endpoints of pubertal alteration and are sufficient to draw conclusions about whether or not a compound should undergo Tier 2 testing. Therefore the organ weights are a time consuming and problematic component that are not needed and could be eliminated. Uterine and ovarian weights are also confounded by cycle. Although the protocol states that the estrous cycle at the time of necropsy should be taken into account, guidance as to how to do this is insufficient.

c. Observations and Measurements

The data sheet for observations and measurements is sufficient and specific details about reservations regarding data collection were given above.

d. Data Compilation

Section XII of the protocol states that uterus, thyroid, ovary and kidney are to be evaluated for pathologic abnormalities but no guidance is given as to how the histological findings are to be quantitatively assessed. It is unclear how the histological data will contribute to the interpretation of the data.

e. Reporting of Results

Data reporting is thorough and acceptable with the exception of the histological data.

5. Strengths/Limitations of the Assay

As discussed in detail above, the assay maximizes breadth at the expense of a robust experimental design to answer a specific question. Removal of the thyroid endpoints and the organ weights would allow for the inclusion of a needed and critical positive control group and make the data vastly easier to interpret. Simplification would also make the assay easier to replicate across laboratories. To assess pubertal disruption, only age at vaginal opening, day of first estrus, and regularity of the estrus cycle (at least 6 weeks post-puberty) is required. All other measures are extraneous and unnecessary. All compounds testing positive in this screening could then be advanced to a Tier-2 screening protocol. If the thyroid elements remain in the protocol, the use of a positive control for the pubertal endpoints is strongly recommended.

6. Impacts

a. Test Substances

Given that no compound to date has tested negative in this assay it is difficult to evaluate the potential effects of test substance on outcome.

b. Analytical Methods

Analysis of variance is to be used with body weight at weaning as the covariate. Day of the estrus cycle is not taken into consideration in the analysis of organ weights but should be for the ovary and uterus as it likely has a greater effect on the weights of these organs than overall body weight.

c. Statistical Methods

See above

7. Repeatability and Reproducibility

Overall the repeatability and reproducibility of the assay is good and has been sufficiently demonstrated. However, the report identifies a number of areas where the data are inconsistent. In nearly all cases, the inconsistencies are noted in endpoints that are not needed to identify the disruption of puberty. Ovarian and uterine weights are unreliable measures of endocrine disruption, particularly when collected without regard to cycle. More guidance as to how cycle should be considered when collecting organ weight is needed.

Inconsistency within the data may also result from variation in animal stress. Without sufficient controls to minimize stress it is likely that inter-laboratory variability will continue to be a problem particularly when large batches of animals have to be sacrificed on a single day and there is thus a lot of human activity in the vivarium. Great care is taken to consider variation in body weight when interpreting the data. In contrast, relatively little information is given to as to how stress was minimized or how housing conditions differed between laboratories. This is a considerable problem. Housing of other animals in the facility, particularly non-human primates, dogs, or cats can increase rodent stress substantially. Transport stress is also a concern. Pups born to dams delivered “timed pregnant” will experience more gestational stress than those born

to mothers impregnated in-house. This may also affect inter-laboratory data collection and consistency. Housing conditions, including diet and day of cage changes, should be matched as closely as possible between laboratories.

References

Alworth, L. C., et al., 2002. Uterine responsiveness to estradiol and DNA methylation are altered by fetal exposure to diethylstilbestrol and methoxychlor in CD-1 mice: effects of low versus high doses. Toxicol Appl Pharmacol. 183**,** 10-22.

Gallo, D., et al., 1999. Reproductive effects of dietary soy in female Wistar rats. Food Chem Toxicol. 37**,** 493-502.

Gioiosa, L., et al., 2007. Developmental exposure to low-dose estrogenic endocrine disruptors alters sex differences in exploration and emotional responses in mice. Horm Behav. 52**,** 307-16.

Goodman, J. E., et al., 2006. An updated weight of the evidence evaluation of reproductive and developmental effects of low doses of bisphenol A. Critical Reviews in Toxicology. 36**,** 387-457.

Kato, H., et al., 2003. Changes in reproductive organs of female rats treated with bisphenol A during the neonatal period. Reprod Toxicol. 17**,** 283-8.

Marcondes, F. K., et al., 2002. Determination of the estrous cycle phases of rats: some helpful considerations. Braz J Biol. 62**,** 609-14.

Patisaul, H. B., Polston, E. K., 2007. Influence of endocrine active compounds on the developing rodent brain. Brain Res Rev.

Rubin, B. S., et al., 2006. Evidence of altered brain sexual differentiation in mice exposed perinatally to low, environmentally relevant levels of bisphenol A. Endocrinology. 147**,** 3681-91.

Rubin, B. S., et al., 2001. Perinatal exposure to low doses of bisphenol A affects body weight, patterns of estrous cyclicity, and plasma LH levels. Environ Health Perspect. 109**,** 675-80.

vom Saal, F. S., 2006. Bisphenol A eliminates brain and behavior sex dimorphisms in mice: how low can you go? Endocrinology. 147**,** 3679-80.

vom Saal, F. S., Hughes, C., 2005. An extensive new literature concerning low-dose effects of bisphenol A shows the need for a new risk assessment. Environ Health Perspect. 113**,** 926-33.

Whitten, P. L., et al., 1993. A phytoestrogen diet induces the premature anovulatory syndorme in lactationally exposed female rats. Biology of Reproduction. 49**,** 1117-1121.

Yener, T., et al., 2007. Determination of oestrous cycle of the rats by direct examination: how reliable? Anat Histol Embryol. 36**,** 75-7.

3.5 Deodutta Roy Review Comments

Review of the Integrated Summary Report (ISR) and Pubertal Protocol

1. Clarity of the stated purpose of the assay in the Integrated Summary Report (ISR) and pubertal protocol (Appendix 1): Though the intent of pubertal assay described in both ISR and the Appendix 1 female pubertal protocol is the same, the description of the purpose of the assay is not exactly the same. For example,

(i) In the Appendix 1 protocol, Section I. Purpose and Applicability, the first sentence states that "the purpose of this protocol is to quantify the effects of chemicals on pubertal development and thyroid function in the intact juvenile/peripubertal female rat."

On page 8, para 3 of ISR, Section III: Purpose of the assay, first sentence: It states that "The purpose of the female 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 endocrine system."

(ii) The Appendix 1 protocol in the section I. Purpose and Applicability, second sentence: It states that "this assay detects chemicals that display antithyroid, estrogenic, or antiestrogenic activity (e.g., alterations in receptor binding or steroidogenesis), or alter hypothalamic function or gonadotropin or prolactin secretion".

On page 8, para 3 of ISR, section III Purpose of the assay, second sentence: It states that "This assay is capable of detecting chemicals with antithyroid, estrogenic, or antiestrogenic activity or agents which alter pubertal development via changes in luteinizing hormone, follicle stimulating hormone, prolactin or growth hormone levels or via alterations in hypothalamic function.

Page 3 and the Table 1 recommended by the Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC) of ISR: Pubertal female (rat): "An assay to detect chemicals that act on estrogen or through the hypothalamus-pituitary gonadal (HPG) axis that controls the estrogen and androgen systems. It is also enhanced to detect chemicals that interfere with thyroid system".

(iii) On a similar smaller note, the ISR title is “Validation of a Test Method for Assessment of on Pubertal Development and Thyroid Function in Juvenile Female Rats as a Potential Screen in The Endocrine Disruptor Screening Program Tier 1 Battery”. In this title the context of use of the report is very clear, i.e., validation. However, in the pubertal assay protocol of the Appendix 1, the title starts from “Test Method----.” The context in which this pubertal protocol would be used was not clear in the title. Moreover, the title of both ISR and appendix 1 should be“------Intact Juvenile/Peripubertal Female Rats” instead of “----Juvenile Female Rats”. This will be consistent with the description in the purpose of the assay and also removes the ambiguity of intact versus ovariectomized.

The above different descriptions in the female pubertal protocol and in the ISR report of the purpose of female pubertal rat assay clearly create ambiguity and confusion, and it does not take into the account new knowledge in the field of endocrine disruptors.

For example, patterns of gonadotropin secretion during puberty in girls have become clearer as measurement techniques have improved. It is now widely recognized that endocrine or paracrine factors different from gonadotropins may play a relevant role as modulators of estrogen (E2) secretion early in the process of ovarian maturation that leads to premature sexual development in girls. A variety of growth factors, including IGF-I, are considered to have a synergistic effect on gonadotropin-induced stimulation of ovarian steroid synthesis or aromatization and breast development.

The purpose of the assay in ISR needs to be re-worded to remove the ambiguity and to make it comprehensive and clear. Here is a draft of an attempt to re-word it: “The purpose of the female pubertal assay is to assess the potential of a chemical substance or mixture to interact with endocrine system which influences pubertal development and thyroid function in the intact juvenile/peripubertal female rat. This assay measures indices of pubertal development and is capable of detecting chemicals interacting with the estrogen, androgen, and thyroid hormonal systems, or agents which alter pubertal development via changes in luteinizing hormone, follicle stimulating hormone, prolactin or growth hormone levels or via alterations in hypothalamic function.”

(iv) Table 3 on page 9 in the Section III: Purpose of the assay lists the end points for the female pubertal protocol: It is not clear why one of the highly sensitive hormone dependent organs, i.e., mammary gland is not included for the analysis of its weights and histopathology. In the various strain of rats, it has been shown that the treatment of 14-21 days with endocrine disruptors, particularly estrogenic in nature produces profound changes in the mammary gland. The central nervous system is one of other system should have been included for analysis of its weight and histopathology, because we now that endocrine disruptors influences its development and functions. Why the levels of estrogen, androgen and progesterone were not proposed to be measured is not clear. It is the ratio of androgen and estrogen or estrogen and progesterone which determines their effects on the target organs.

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

In the pubertal protocol (Appendix 1), Table 6 entitled "Potential changes indicative of different mode of action that may be observed in female pubertal protocol" provides very clear comprehensive summary of expected different effects that may be observed from different modes of action. Using this table, chemical substances that that exert effects via various mechanisms or different modes of interaction with the endocrine system can be identified. The description of the data interpretation is very consistent with the stated purpose of the assay. The guidelines described in the text given for data interpretation for doses level tested, explanation of negative results in the context of interaction with endocrine system, performance criteria, and evaluations of endpoints are very clear. The same is true for ISR report which describes this in the data interpretation section (page 74-77).

3. Biological and toxicological relevance of the assay as related to its stated purpose: The pubertal period is a very sensitive age for exposure to agents which alter the endocrine system. Therefore, this assay when validated should be able to detect chemical disruptors of estrogen, androgen, and thyroid action. The female pubertal rat assays can also identify compounds that alter hypothalamic-pituitary control of the gonads or thyroids.

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

The protocol is well written. . Examples of Tables 1-5 must be very helpful for the contract laboratories in measuring endpoints and preparing reports. It is assumed that all studies were conducted in professional contract laboratories with GLP facility. On minor points, it was not clear from the document whether each contractor purchased the animal, chemicals, and kits from the same source and the protocol of each laboratory assay was the same. This would have reduced some of the variability. The descriptive text in the protocol needs some further improvement for the clarity. As described above, the first sentence of the protocol states that "the purpose of this protocol is to quantify the effects of chemicals on pubertal development and thyroid function in the intact juvenile/peripubertal female rat." The word "quantify the effects---" is misleading because some of the data, particularly histology of thyroid section are qualititative and semi-quantitative in nature. It would be appropriate to use "determine or investigate the effects----" instead of "quantify the effects-----."

The second sentence of this section states that "this assay detects chemicals that display antithyroid, estrogenic, or antiestrogenic activity (e.g., alterations in receptor binding or steroidogenesis), or alter hypothalamic function or gonadotropin or prolactin secretion". Are these measures or effects of chemicals described in the second sentence indices of pubertal development is not clear? The purpose of the protocol should clearly match with assay objectives with multiple endpoints. The impaired pubertal development includes early or delayed onset of puberty, impaired gonadal maturation (steroidogenesis) or ovarian function, shown by decreased or increased estrogen levels, impaired secretion of gonadotropin, thyroid hormones or prolactin. The clear connection between first two sentences is missing. The second sentence should be re-worded as "The assay is expected to identify the endocrine-mediated effects on female pubertal

development by measuring puberty indices following exposure to chemicals with estrogenic or anti-estrogenic activity, inhibitors of steroid and thyroid hormone synthesis,-----."

It is not clear how much blood is needed for hormone assay and at what speed it should be centrifuged. For the methodology of hormone measurement, four methods are described and it appears that the choice to choose was left to the contract laboratory. If available, the preferred choice of assay should have been time-resolved immunofluorometric assays (IFMA) particularly for measurement of gonadotropin concentrations. This is more sensitive than radioimmunoassay or immunoradiometric assays. This would have helped in reducing intra-laboratory variability

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

The IRS report very clearly describes various studies of low and high doses of chemical substances conducted in the different laboratories. Each contract laboratory showed that the pubertal assays can identify compounds that alter hypothalamic-pituitary control of the gonads. All studies using thyroid-active agents showed that the female pubertal assays detect alterations in thyroid function following exposure to compounds interacting to thyroid system. I concur with EPA conclusion in regards to the strengths and weakness of various assays. The one of the major strengths of this study is that it is an in vivo assay, and it can measure the effects of both parent compounds and their metabolites. This assay estimates the interaction with the endocrine system. Additionally, this assay measures the effects of the endocrine disruptors at one of the critical time period of the development of the animal, which is highly sensitive to changes in the endocrine system. This would help in identifying weak endocrine disruptors. Use of the redundant multiple endpoints increased the credence of the assay. This was further strengthened by the use of very well though performance criteria.

Minor weakness

It has been shown that the rodent pubertal female assay is useful for identifying potential endocrine disruptors having not only estrogenic/antiestrogenic but also

androgenic/antiandrogenic activities, therefore it is not clear why androgenic/antiandrogenic activities were not monitored.

We now know the non-receptor-mediated mechanisms exist by which unknown disruptors can affect the embryo/fetus without showing positive effects on the proposed classical multiple endpoints. In situ biochemical and gene activation measurements or biomarkers for assessing pubertal development could have really helped to detect subtle changes in the endocrine systems which would be not detected otherwise by proposed multiple endpoints in this assay. ChIP on ChIP assay would have been more sensitive for screening effects of endocrine disruptors by studying changes in genes involved in androgen, estrogen, or thyroid systems.

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 choice of test substances and analytical methods is well thought and is well described in this document. It has been identified in the report that the cost associated with animal experiments did not allow to use appropriate ranges of positive and negative test substances, particularly weak positive controls. However, the use of weak positive compound was critical for validation of this study because most of the unknown test compounds are hormonally weak substances.

The methodologies for measuring indices of puberty are not very modern, and they may not be very sensitive in detecting the initiation and progression of molecular changes that ultimately impair the pubertal development. There is a concern that all these functional assays may not be able to detect subtle changes in the animals exposed to weak endocrine disruptors. The animal strains used are not the most sensitive to estrogenic compounds, which makes it further difficult in judging the suitability of analytical methods.

Statistical methodology is not my expertise, so I have decided not to make any comment regarding this.

7. Repeatability and reproducibility of the results obtained with the assay:

Based on the ISR document, it appears that results obtained with the pubertal assay in the different contract laboratories are repeatable and reproducible, because all three laboratories data showed that the female pubertal assay may be useful for identifying chemicals that operate through a variety of mechanisms. This was true for both estrogenic and thyroid system interacting chemicals. For example, the TherImmune 1 study used a single dose of six different compounds in three different laboratories. Three different laboratories identified expected endocrine effects from exposure to chemicals with estrogenic, anti-estrogenic, androgenic or anti-androgenic activity, inhibitors of steroid and thyroid hormone synthesis, and a dopamine antagonist. ethynyl estradiol, tamoxifen (e.g, antagonist and partial estrogen agonist), and methoxychlor advanced the onset of vaginal opening. Propylthiouracil (e.g., an inhibitor of thyroid hormone synthesis), ketoconazole (e.g., an inhibitor of steroid synthesis) or pimozide (e.g., a dopamine antagonist) delayed the age of vaginal opening. The sensitivity of the protocol was assessed through multi-chemical study. Two different doses of six compounds were used for this study. The low doses of all six compounds showed expected changes in the estrogen-related and thyroid system-related endpoints. The multi-dose study (TherImmune1 2) used three compounds, ethynyl estradiol, methoxychlor and phenobarbital and showed similar sensitivity to estrogenic compounds. Thus, the EPA in this ISR document has very correctly concluded that the female pubertal protocol is transferable, sensitive and reproducible.

Appendix A

CHARGE TO PEER REVIEWERS

PEER REVIEW CHARGES

for

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

October 1, 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 female 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 endocrine system. This assay is capable of detecting chemicals with antithyroid, estrogenic, or antiestrogenic [estrogen receptor (ER) 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 22 - 42 (21 days) to 15 females per dose level. The endpoints are growth (body weight); age and weight at vaginal opening; organ weights (uterus, blotted; ovaries; thyroid; liver; kidneys; pituitary; adrenals); histology of uterus, ovary, thyroid, kidney; serum thyroxine, total; serum thyroid stimulating hormone; age at first estrus, length of estrous cycle, percent of animals cycling, and percent of animals cycling regularly.

Although peer review of the female 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 female pubertal rat assay, especially the inter-laboratory validation exercise. Review and comment shall be directed to each of the following:

1. Clarity of the stated purpose of the assay.

2. Clarity, comprehensiveness and consistency of the data interpretation 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 Validation of a Test Method for Assessment of Pubertal Development and Thyroid Function in Juvenile Female Rats as a Potential Screen in the Endocrine Disruptor Screening Program Tier-1 Battery (PDF) (122pp, 351K)

Appendix C

SUPPORTING MATERIALS

Appendix 1. Female Pubertal Protocol

Protocol for the Female Pubertal Rat Assay (PDF) (19pp, 91K)

Explanation of the 5 Thyroid Slides (attached below) and How they are Used in the Protocol for the Female 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 Female Pubertal Rat Assay Endocrine-Disrupting Chemicals: Prepubertal Exposures and Effects on Sexual Maturation and Thyroid Activity in the Female 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:

Goldman, J.M., Laws, S.C., Balchak, S.K., Cooper, R.L., Kavlock, R.J.. (2000). Endocrine-Disrupting Chemicals: Prepubertal Exposures and Effects on Sexual Maturation and Thyroid Activity in the Female Rat. A Focus on the EDSTAC Recommendations. Crit Rev.Toxicol. 30(2), 135-196.

Appendix 3. Transferability Study (TherImmune 1) Summary Report

Assessment of Pubertal Development and Thyroid Function in Juvenile Female Rats (Study 101) (PDF) (184pp, 4.7M)

Assessment of Pubertal Development and Thyroid Function in Juvenile Female Rats (Study 103) (PDF) (184pp, 4.8M)

Appendix 4. 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 Female Rats (XLS) (5pp, 179K)

Appendix 5. Multi-Chemical Study (RTI) Summary Report

Final Report from the RTI Multi-Chemical Study - Assessment of Pubertal Development and Thyroid Function in Juvenile Female CD® (Sprague-Dawley) Rats After Exposure to Selected Chemicals Administered by Gavage on Postnatal Days 22 to 42/43 (PDF) (182pp, 543K)

Pathology Report from the RTI Multi-Chemical Study - Assessment of Pubertal Development and Thyroid Function in Juvenile Female CD® (Sprague-Dawley) Rats After Exposure to Selected Chemicals Administered by Gavage on Postnatal Days 22 to 42/43 (PDF) (52pp, 16.9M)

Feed Analysis Reports from the RTI Multi-Chemical Study - Assessment of Pubertal Development and Thyroid Function in Juvenile Female CD® (Sprague-Dawley) Rats After Exposure to Selected Chemicals Administered by Gavage on Postnatal Days 22 to 42/43 (PDF) (5pp, 239K)

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 Female CD® (Sprague-Dawley) Rats After Exposure to Selected Chemicals Administered by Gavage on Postnatal Days 22 to 42/43 (PDF) (1pp, 8K)

Appendix 6. 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 Female CD® (Sprague-Dawley) Rats After Exposure to Selected Chemicals Administered by Gavage on Postnatal Days 22 to 42/43 (XLS) (11pp, 115K)

Notes for Appendix 6 - Multi-chemical study - RTI - Summary Table (PDF) (4pp, 16K)

Appendix 7. 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 Female CD® (Sprague-Dawley) Rats After Exposure to Selected Chemicals Administered by Gavage on Postnatal Days 22 to 42/43 (PDF) (33pp, 48K)

Appendix 8. Multi-Dose Study (TherImmune 2) Summary Report

Final Report - 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 9. 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) (16pp, 74K)

Notes for Appendix 9 - Multi-dose study - TherImmune 2 - Summary Table (PDF) (4pp, 20K)

Appendix 10. 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) (4pp, 58K)

Appendix 11. White Paper on Rat Strain Differences

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

Appendix 12. 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 13. Interlaboratory Validation Study Summary Report (Charles River/Argus)

Summary Report for the CR-DDS Argus Division's Interlaboratory Validation of the Female Pubertal Assay (PDF) (38 pp., 311KB)

Appendix 14. Interlaboratory Validation Study Summary Report (Huntingdon)

Summary Report for the Huntingdon Life Science's Interlaboratory Validation of the Female Pubertal Assay (PDF) (51 pp, 1.9M)

Appendix 15. Interlaboratory Validation Study Summary Report (WIL)

Summary Report for the WIL Research Laboratories' Interlaboratory Validation of the Female Pubertal Assay (PDF) (53 pp, 850K)

Appendix 16. Interlaboratory Validation Study Analysis Report (Battelle)

Battelle Report on the Analysis of the Interlaboratory Validation of the Female Pubertal Assay Assessment of Pubertal Development and Thyroid Function in Juvenile Female Rats (PDF) (225pp, 7M)

Appendix 17. 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 Female Pubertal Assay Assessment of Pubertal Development and Thyroid Function in Juvenile Female Rats (XLS) (13pp, 148K)

Appendix 18. Interlaboratory Validation Study Comparison of Results Table

Table with the Comparison of Results from the Female Pubertal Interlaboratory Validation Study (PDF) (10pp, 81K)