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

03-Aug-2009

English - Or. English

ENV/JM/MONO(2009)32
Unclassified

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

SERIES ON TESTING AND ASSESSMENT
Number 110

REPORT OF THE VALIDATION PEER REVIEW FOR THE WEANLING HERSHBERGER BIOASSAY AND AGREEMENT OF THE WORKING GROUP OF NATIONAL COORDINATORS OF THE TEST GUIDELINES PROGRAMME ON THE FOLLOW-UP OF THIS REPORT

JT03268419

English - Or. English

OECD Environment, Health and Safety Publications

Series on Testing and Assessment

No. 110

REPORT OF THE VALIDATION PEER REVIEW FOR THE WEANLING HERSHBERGER BIOASSAY AND AGREEMENT OF THE WORKING GROUP OF NATIONAL COORDINATORS OF THE TEST GUIDELINES PROGRAMME ON THE FOLLOW-UP OF THIS REPORT

IOMC

INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS

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

Environment Directorate

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

Paris 2009

Also published in the Series on Testing and Assessment:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

No. 39, *Guidance Document on Acute Inhalation Toxicity Testing (2009)*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Multiple Laboratories. Surgical Castrate Model Protocol (2007)

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

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

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

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

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

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

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

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

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

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

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

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

No. 86, *Report of the OECD Validation of the Rodent Hershberger Bioassay: Phase 2: Testing of Androgen Agonists, Androgen Antagonists and a 5 α-Reductase Inhibitor in Dose Response Studies by Multiple Laboratories (2008)*

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

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

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

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

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

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

No.93 . *Report of the Validation of an Enhancement of OECD TG 211: Daphnia Magna Reproduction Test (2008)*

No.94. *Report of the Validation Peer Review for the 21-Day Fish Endocrine Screening Assay and Agreement of the Working Group of the National Coordinators of the Test Guidelines Programme on the Follow-up of this Report (2008)*

No.95 *Detailed Review Paper on Fish Life-Cycle Tests (2008)*

No.96 *Guidance Document on Magnitude of Pesticide Residues in Processed Commodities (2008)*

No.97 *Detailed Review Paper on the use of Metabolising Systems for In Vitro Testing of Endocrine Disruptors (2008)*

No. 98 *Considerations Regarding Applicability of the Guidance on Transformation/Dissolution of Metals Compounds in Aqueous Media (Transformation/Dissolution Protocol) (2008)*

No. 99 *Comparison between OECD Test Guidelines and ISO Standards in the Areas of Ecotoxicology and Health Effects (2008)*

No.100 *Report of the Second Survey on Available Omics Tools (2009)*

No.101 *Report on the Workshop on Structural Alerts for the OECD (Q)SAR Application Toolbox (2009)*

No.102 *Guidance Document for using the OECD (Q)SAR Application Toolbox to Develop Chemical Categories According to the OECD Guidance on Grouping of Chemicals (2009)*

No.103 *Detailed Review Paper on Transgenic Rodent Mutation Assays (2009)*

No.104 *Performance Assessment: Conparsion of 403 and CxT Protocols via Simulation and for Selected Real Data Sets (2009)*

No. 105 *Report on Biostatistical Performance Assessment of the draft TG 436 Acute Toxic Class Testing Method for Acute Inhalation Toxicity (2009)*

No.106 *Guidence Document for Histologic Evalution of Endorcine and Reproductive Test in Rodents (2009)*

No.107 *Preservative treated wood to the environment for wood held in storage after treatment and for wooden commodities that are not cover and are not in contact with ground.(2009)*

No.108, *Intact, Stimulated, Weanling Male Rat Version of the Hershberger Bioassay(2009)*

No.109, *Literature Review on the 21-Day Fish Assay and the Fish Short-Term Reproduction Assay (2009)*

No.110, *Report of the Validation Peer Review for the Weanling Hershberger Bioassay and Agreement of The Working Group of National Coordinators of the Test Guidelines Programme on the Follow-Up of this Report (2009)*

ABOUT THE OECD

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

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

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

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

**OECD Environment Directorate,
Environment, Health and Safety Division**

**2 rue André-Pascal
75775 Paris Cedex 16
France**

**Fax: (33-1) 44 30 61 80
E-mail: ehscont@oecd.org**

FOREWORD

This document presents the peer review report (PRP) for the validation of the Weanling Hershberger Assay, preceded by the agreement of the Working Group of National Coordinators of the Test Guidelines Programme (WNT) on the follow-up of the PRP report.

The peer review was sponsored by United States (Environmental Protection Agency). In June 2008, the National Coordinators were requested to nominate candidate peer reviewers.

This document is published on the responsibility of the Joint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the Organisation or the governments of its member countries.

This document is published on the responsibility of the Joint Meeting of the Chemicals Group and Management Committee of the Special Programme on the Control of Chemicals of the OECD.

Contact for further details:
Environment, Health and Safety Division
Environment Directorate
Organisation for Economic Co-Operation and Development
2, rue André Pascal
75775 Paris Cedex 16, France

Tel : 33-1-45-24-16-74
E.mail : env.edcontact@oecd.org

Agreement of the Working Group of the National Coordinators of the Test Guidelines Programme on the Follow-up of the Peer Review Report

The peer review report of the validation of the weanling version of the Hershberger assay was submitted to the Working Group of National Coordinators of the Test Guidelines Programme (WNT) on 27 January 2009, for information. This report is complementary to the peer review report of the validation of the castrate version of the Hershberger assay [N° 85 in the Series on Testing and Assessment].

The WNT noted that the weanling version of the Hershberger Bioassay did not appear to be able to consistently detect effects on androgen-dependent organ weights from weak anti-androgens at the doses tested in the validation studies. The WNT also noted that the peer review panel was divided on the issue of the inclusion of the intact weanling model in the draft Test Guideline.

Considering that the weanling model is less responsive than the castrate model when administering weak anti-androgens, the WNT went back on its 2007 decision and agreed that the intact weanling model should not be included in the Test Guideline, while the weanling model will be included in a Guidance Document.

REPORT OF THE VALIDATION PEER REVIEW FOR THE WEANLING HERSHBERGER BIOASSAY

INTRODUCTION

The purpose of this work assignment, sponsored by the U.S. Environmental Protection Agency, was to administer an independent peer review of a report summarizing the Organization for Economic Cooperation and Development (OECD) validation of the weanling rat model as an alternative to the castrate adult male rat that is traditionally used in the Hershberger Assay. The peer review was administered by Battelle in December 2008 and January 2009. The subject of the peer review was the following report:

> ***Intact, Stimulated, Weanling Male Rat Version of the Hershberger Bioassay, Task Order 16: Immature Male Rat Hershberger Model Validation Report, EPA Contract No. EP-W-06-026, Project No. 0210114.016, prepared for U.S. Environmental Protection Agency, Endocrine Disruptor Screening Program; prepared by RTI International, Research Triangle Park, NC, October 27, 2008.***

The peer review panel members were also provided with the following supporting document for reference:

> ***Draft OECD Guideline for the Testing of Chemicals. The Hershberger Bioassay in Rats: A Short-term Screening Assay for (Anti)Androgenic Properties, Revised 20 October 2008.***

A panel of five peer reviewers were selected based on a list of potential reviewers provided by EPA, supplemented by an independent search conducted by Battelle. Potential reviewers were screened as to their expertise, availability, and lack of real or perceived conflicts of interest. Relevant expertise included the following disciplines: mammalian reproductive or developmental toxicology, andrology/male reproductive function, endocrinology, and/or sexual differentiation and development. Qualifications of each candidate reviewer were submitted to EPA for concurrence and approval. Each reviewer completed a Certification Concerning Conflicts of Interest, verifying that no actual or potential conflicts exist. The members of the peer review panel and their affiliations are as follows:

- **Ibrahim Chahoud, DVM, PhD,** Department of Reproductive Toxicology, Institute of Clinical Pharmacology and Toxicology, Department of Toxicology, Charitè University Medical School, Campus Benjamin Franklin, Berlin, Germany.

- **John Charles Eldridge, PhD,** Professor of Physiology and Pharmacology, Wake Forest University School of Medicine, Department of Physiology and Pharmacology, Winston-Salem, North Carolina, USA.

- **Heather B. Patisaul, PhD,** Assistant Professor, Department of Biology; Associate Member, Department of Toxicology, North Carolina State University, Raleigh, North Carolina, USA.

- **Richard M. Sharpe, PhD,** Special (Professorial level) Appointment, Medical Research Council Human Reproductive Sciences Unit, Programme leader: Physiological basis for male reproductive development and function, Professor, College of Medicine & Veterinary Medicine, University of Edinburgh, Edinburgh, Scotland, United Kingdom.

- **Dr. Chris E. Talsness, DVM,** Charité Universitätsmedizin Berlin, Institute of Clinical Pharmacology and Toxicology, Working Group Reproductive Toxicology, Campus Benjamin Franklin, Berlin, Germany.

A kickoff conference call was held with the EPA, Battelle, and all five reviewers on January 12, 2009, for the purpose of providing background, outlining the scope of work, and resolving any questions or concerns on the part of the reviewers. Beyond that, all reviewers worked independently of each other. The peer reviewers were asked to read the reports cited above and to respond to a series of charge questions prepared by EPA.

The comments from the peer reviewers specific to each charge question are compiled below, followed by any general or summary comments provided by the peer reviewers. The language from each charge question is presented before the reviewers' responses. The original, verbatim submissions from the peer reviewers are included in Appendix A. A separate Peer Review Record is also being prepared for submission to EPA, with further documentation of the reviewers' qualifications and other supporting material.

OVERVIEW OF REVIEWER CONCLUSIONS

To summarize across all reviewers, the following final conclusions were submitted:

- The substitution of the traditional castrate version of the Hershberger assay with the weanling version **is justified.**

- The weanling version of the Hershberger assay appears to be a **barely adequate** substitute for the traditional castrate version.

- Therefore the weanling version of the Hershberger bioassay is **not an adequate substitute** for the traditional castrate version.

- The intact weanling assay is fit for purpose and can provide **an adequate substitute** for the castrated rat assay.

- The weanling version **could be used as a substitute** for the castrated version with the caveat that higher doses may have to be employed. The choice of which version to employ could be based on available *in vitro* data.

Reviewer Responses to Charge Questions

1. Comment on the adequacy of the validation program by stating how well it meets the following validation criteria:

A. The rationale for the test method should be available.
This should include a clear statement of the scientific basis, regulatory purpose and need for the test.

Dr. Chahoud. *Scientific basis:* The scientific basis for the Hershberger Assay is well documented and employed for decades since the 1930's. The aim of the test is the evaluation of the ability of chemicals to change the weight of androgen-dependent tissues in prepubertal male rats. The test period lasts from approximately PND 21-33. At this age the tissues evaluated in male rats exhibit androgen receptors and at the same time serum testosterone concentrations are low. Therefore, the androgen dependent tissues are sensitive to exogenous androgens as well as anti-androgens. In contrast to the castrated version, the advantage of this model is the intact hypothalamic-pituitary-gonadal axis which provides a physiologically normal test situation.

Regulatory purpose: The regulatory need exists to rapidly assess and evaluate a chemical as a possible androgen agonist or antagonist or 5α-reductase inhibitor. The Hershberger bioassay serves as a mechanistic *in vivo* screening assay and its application should be seen in the context of the "OECD Conceptual Framework for the Testing and Assessment of Endocrine Disrupting Chemicals" (annex 2). The assay is part of Level 3 and is designed to provide data about a single endocrine mechanism, i.e. (anti)androgenicity. The inclusion in a battery of *in vitro* and *in vivo* tests to identify substances with potential to interact with the endocrine system is recommended, ultimately leading to hazard and risk assessments for human health or the environment.

Need for the test: Whether a substance acts as androgen or antiandrogen can be assessed in an *in vitro* test battery. However, the disadvantage of *in vitro* tests is their lack of information about the toxicokinetic properties of a given substance. Therefore, there is a need for *in vivo* tests in order to be able to get important information about resorption, metabolisation and elimination of the tested compound. The Hershberger Assay is an adequate test system for these purposes. Evaluation of the test as an alternative protocol to the castrated model version is justified.

Dr. Eldridge. All of the test methods and the scientific basis for the methods were clearly presented. The needs for this test (Hershberger), the justification underlying the need to investigate the present modifications (use of intact weanling rats), and the regulatory purposes were also clearly presented.

Dr. Patisaul. The Immature Male Rat Hershberger Model Validation Report (dated November 30, 2008) clearly states that the "overall aim of the validation program is to demonstrate that the Hershberger bioassay is a robust, sensitive, reliable and reproducible bioassay that can be considered as the basis for an OECD Test Guideline. Once available, the test guideline is intended to be used as one element in an overall testing strategy for the detection and assessment of potential endocrine disruptors." Use of the Hershberger bioassay as a component of an endocrine disruptor screening program is appropriate and necessary.

The primary rationale for recommending the weanling version of the Hershberger Assay, instead of the castrate version, is the desire to avoid castration for animal welfare reasons. Another rationale for the switch is stated in the final paragraph of the report. Because the

weanlings are gonadally intact, and therefore the hypothalamic-pituitary-gonadal (HPG) axis is theoretically capable of responding to the administration of an endocrine active compound, the weanling version of the Hershberger bioassay is potentially advantageous over the castrate version because it has the potential to detect effects resulting from disruption within the (HPG) axis in addition to direct action on androgen receptors. However, the high level of TP needed for the weanling version (1 mg/kg compared to 0.4 for the castrate version) would likely suppresses the responsivity of the HPG axis through steroid negative feedback. Therefore the ability of the HPG axis to adequately respond to the chemical insult under the test conditions of the Hershberger bioassay is likely minimal.

Another stated rationale for using the weanling version of the Hershberger bioassay over the castrate version is concern that the results obtained using the castrate version might not be extrapolable to gonad-intact animals. It is unlikely that either version would produce results that are unequivocally extrapolable to the uncastrated adult (or juvenile), but this should not be considered a major problem because the primary goal of the assay is to screen for compounds with endocrine disrupting properties, not to make predictions about how the compound might affect gonadally intact individuals. For the purposes of screening, the assay with the greatest sensitivity, higher demonstrated degree of reproducibility, and lowest laboratory intervariability should be considered superior.

Dr. Sharpe. The rationale for the assay, its mechanistic/functional basis and the specific reasons why this additional test is being evaluated are well described, as are specific issues that require consideration in the validation and evaluation exercise.

Dr. Talsness. The *scientific basis* for the test is founded and has been well documented and employed in assays since the 1930's. In addition, extensive validation work has successively been performed on the castrated version of the assay. The test evaluates the ability of a substance to increase or decrease the weight of androgen-dependent tissues of the pre-pubertal male. At the time of the test, starting between PND 21-24 and ending between PND 30-33, androgen receptors are present on the tissues evaluated and the serum concentration of testosterone is relatively low and begins to slowly increase over time with a peak occurring between 50 and 60 days of age. Due to the relatively low testosterone concentrations, the androgen dependent tissues can respond to exogenous androgens and anti-androgen effects can be demonstrated with co-administration of testosterone propionate in comparison with testosterone propionate alone. 5α reductase inhibitors can be theoretically detected based on a differentiated response among the tissues due to their differing relative dependency on 5 α reductase conversion, although validation of this was not performed, e.g., finasteride. The hypothalamic-pituitary-gonadal axis is functional in this model and indirect effects on this axis can theoretically be detected although validation for this was not performed, e.g., GnRH agonists or antagonists.

The *regulatory purpose* of this assay is to serve as a Level 3 test which includes "*in vivo* assays providing data about single endocrine mechanisms and effects". Although the weanling assay theoretically covers a larger scope of investigation than the castrated version, only information regarding (anti) androgenicity in the broad sense is obtained. Screens should be sensitive methods with the emphasis placed on reducing the number of false negatives while possibly increasing the number of false positives. The weanling version is clearly less sensitive than the castrated version and appears to have similar specificity as the castrated version. The doses required for identification of substances were higher in the weanling version than the castrated one, however, each chemical was eventually identified at the highest dose as in the castrated version, i.e., the resulting information obtained from the two tests was the same. This test guideline, therefore, could be used for hazard characterization. Calculation of reference doses should be performed with data derived from other *in vivo* tests investigating other types of endpoints.

The *need* for the test is to reduce the number of chemicals identified in level 2 testing which will require further investigation using more elaborate protocols to evaluate possible endocrine-related effects and to address welfare concerns questioning the need to use a castrated model to achieve this aim.

In vitro tests do not account for the influences of absorption, distribution, metabolization or elimination (ADME) and it is theoretically possible that "positive compounds" from *in vitro* screens will be rendered "inactive" in the *in vivo* model, thereby lowering the number of chemicals indicated for further evaluation of endocrine effects (level 4). It is just as theoretically possible that negative compounds in the *in vitro* testing phase could be positive *in vivo* due to the influences of ADME which leads to the question whether the Hershberger Bioassay should be reserved instead for chemicals which are negative in level 2.

Evaluation of the test as an alternative protocol to the castrated model version is justified.

B. The relationship between the test method's endpoint(s) and the (biological) phenomenon of interest should be described.
This should include a reference to scientific relevance of the effect(s) measured by the test method in terms of their mechanistic (biological) or empirical (correlative) relationship to the specific type of effect/toxicity of interest. Although the relationship may be mechanistic or correlative, test methods with biological relevance to the effect/toxicity being evaluated are preferred.

Dr. Chahoud. The aim of the weanling version of the Hershberger Assay is the identification of androgen receptor agonists/antagonist and 5α-reductase inhibitors. The endpoints required are weight changes in androgen-dependent tissues (Cowper's glands, levator ani-bulbocavernosus muscle complex, seminal vesicles with coagulating glands and their fluids and ventral prostate). Optionally, other reproductive organs (testes, epididymides) as well as liver, kidney, adrenal glands and hormone concentrations are recommended.

Weight changes in androgen-dependent reproductive tissues are sensitive and relevant markers for androgen agonists and antagonists. Agonists will cause an increase in the weights of the respective tissues, while treatment with antagonists, will prevent the androgen-dependent organ weight increase. The choice of rat accessory sex organs is biologically relevant, mechanistically sensitive and adequate to detect the effects of androgen receptor agonists, antagonists and 5 α-reductase inhibitors.

The Hershberger assay is currently the best available *in vivo* assay for detecting androgen receptor agonists and antagonists. However, it should be stressed that this bioassay can only be used as a screening test.

Dr. Eldridge. The test methods and the chosen parameters appear to be scientifically valid and biologically relevant. The report included a discussion of published literature and contained a good list of citations, both research papers and regulatory reports of assay validation. The targeted tests also appear to be relevant to both human health and to environmental safety of many species of male animals.

Dr. Patisaul. In general, the Immature Male Rat Hershberger Model Validation Report (dated November 30, 2008) clearly and adequately addresses the limitations of the weanling model, compared to the castrate model, the most significant of which is the potential for the

(HPG) axis to respond to the compounds being screened and therefore interfere with the outcome. The most problematic drawback of using gonadally intact weanlings is that the sensitivity of the bioassay is clearly and markedly reduced. To compensate for this, greater numbers of animals would likely be needed to achieve sufficient statistical power which, to me, is a far greater animal welfare concern than castration. Use of an additional assay to validate the results may also be required.

It would also be very difficult to draw definitive conclusions about the potential mechanisms by which a screened compound is producing its effect(s) because the presence of an intact HPG axis allows for numerous alternative pathways other direct action on androgen receptors (ARs). The castrated male Hershberger bioassay provides clear, readily interpretable data on (anti-) androgenicity mediated by (ARs). In the weanling version, additional mechanisms, including metabolic inhibition and hypothalamic and/or pituitary regulation of the gonad, and/or indirect effects on the intact HPG axis, are also possible, making the interpretation of the results more complicated. However, when (anti-)androgen action via a non-AR mechanism is suspected, the weanling version of the Hershberger bioassay may be a more appropriate choice.

Another possible confound of the weanling version of the Hershberger bioassay, which was not addressed in the report, is the potential for the administration of testosterone propionate (TP) to advance pubertal onset. Male rats normally undergo pubertal maturation around 45 days of age at which point gonadotropin secretion elevates to adult levels. This is why it is suggested that animals between 21 and 24 days be used in the gonadally intact weanling version of the assay (so that sacrifice will occur a week before pubertal onset). It is possible that the administration of TP (and or the compounds being screened) will "awaken" the HPG axis and accelerate pubertal onset in the test animals, which would then further reduce the sensitivity of the assay. It is difficult to reliably and accurately determine if a male rat is entering puberty. Prepucial separation is one marker that is often used but this measure is not reliable and could be confounded by any of the compounds used in the assay. Timing of pubertal onset is not a concern in the castrate version of the Hershberger bioassay because the animals are already reproductively mature at the time of castration.

Dr. Sharpe. The reasons for choice of endpoint measurements are well described and sensible and are fully justified, biologically. Performance criteria are established and available and expected working practices and evaluation methods are fully detailed and sound.

Dr. Talsness. The required endpoints include: daily body weight and clinical observations and weights of Cowper's glands, levator ani-bulbocavernosus muscle complex, seminal vesicles with coagulating glands and their fluids and ventral prostate. Optional assessments include daily food consumption, weights of paired testes, paired epididymides, liver, paired kidneys and paired adrenal glands and serum LH, testosterone, T4 and T3.

Evaluation of clinical appearance and daily body weight are reasonable endpoints to provide relevant information to assess possible overt toxicity and allow for adjustment of dose during a period of rapid growth. In addition, body weight allows analyses to be performed regarding possible influences of body weight on weights of accessory sex tissues, the main endpoints of the assay. Change in the weights of the accessory sex tissues is a biologically sound expectation to exposure to exogenous androgens or co-exposure to androgen/anti-androgen because these tissues are dependent on testosterone and 5 α dihydrotestosterone throughout puberty and adulthood for maturation and function. The weanling model is characterized by 1- low baseline serum testosterone concentrations allowing the detection of a response to androgen agonists as the tissues are not in a state of maximal stimulation, 2- the presence of androgen receptors on the accessory sex tissues

allowing a response and 3- steady physiological growth of these tissues with any dramatic changes typically occurring after the timeframe of the study period.

Optional Endpoints: Paired weights of the testes and epididymides

- Androgen agonists

Detection of statistically significant changes in the weights of the epididymides and the testes required a higher dose of TP compared to the mandatory accessory sex tissues. In addition, the relative response of the epididymides is also lower perhaps making detection with this organ more difficult.
Only one laboratory detected changes in the epididymides in response to TREN while the testes data (dose and response) supported the effects observed in the mandatory accessory sex tissues.

- Androgen antagonists

A higher dose of FLU was required to detect a statistically significant response in paired testes weight and statistically significant changes were not observed in response to LIN or DDE.

Epididymides responded at to FLU at the same dose as the mandatory accessory sex tissues and did not respond to LIN or DDE at all doses tested.

In general, the opportunity to evaluate these two organs did not increase the sensitivity of the assay and they proved to be "weak detectors" of androgen antagonists, but changes in weight (although not statistically significant) may be useful as further evidence when equivocal results are obtained with the mandatory accessory sex tissues.

C. A detailed protocol for the test method should be available.
The protocol should be sufficiently detailed and should include, e.g., a description of the materials needed, such as specific cell types or construct or animal species that could be used for the test (if applicable), a description of what is measured and how it is measured, a description of how data will be analyzed, decision criteria for evaluation of data and what are the criteria for acceptable test performance.

Dr. Chahoud. The protocol is clearly written, comprehensive and there are adequate descriptions of the materials, equipment and methodology. Furthermore, performance criteria and justification of animal numbers are acceptable. The description of data analysis is comprehensive in the use of correct statistical approaches.

It is not clear if a statistically significant reduction in only one of the tissues weighed would be sufficient to be called a positive signal or outcome?

Dr. Eldridge. The methods were presented in good detail. It would not be difficult for a professional laboratory to conduct these tests in the manner intended by the methodology designers. Criteria for data evaluation and test performance were also described in detail, and appear to be reasonable.

Dr. Patisaul. In general, the protocol is clearly written and appropriate, however I have a few concerns about some of the details.

1. *DIET:* The diet should be free of phytoestrogens. As written, there are no restrictions on the diet although each lab is supposed to report which diet was used. The presence of phytoestrogens, even in small quantities, needlessly impairs the sensitivity of the assay and introduces inter-laboratory variability. The Draft OECD Guideline for the Testing of Chemicals (Revised October 20, 2008) suggests that (page 8, line 290), "dietary levels of phytoestrogens should not exceed 350 µg of genistein equivalents/gram of laboratory diet." 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 or at least minimize this potentially problematic source of endocrine disrupting compounds.

2. *LITTER EFFECTS:* It is not clear from how many litters the group of weaned males will be pulled. The protocol states that up to 6 weanlings can be housed together but it is not clear if all 6 can be pulled from the same litter or should be combined from different litters. Quality of maternal care can impact male sexual development and therefore the potential for a significant litter effect is a concern that should be mitigated by clearly stating that the males must come from multiple (preferably at least 3) litters.

3. *TIMING OF ADMINISTRATION:* It is not stated at which time of day the compounds are to be administered. Endogenous secretion of androgens has a distinct circadian pattern with levels generally higher in the morning. Therefore the time at which the test compounds are administered should be standardized across labs to ensure that endogenous androgen levels are similar. This is not a major concern in the adult castrate version of the assay because the testes are removed.

Dr. Sharpe. The test requirements in terms of animals, treatment regimes and routes and standard operating procedures for retrieving endpoint tissues for weighing are well described in the relevant *Draft OECD Guideline for the Testing of Chemicals* document. This is important information for this particular assay as variability in consistency of tissue recovery is of fundamental importance in determining the effective working of the assay; some of the endpoint tissues (SVCG, VP, and to a lesser extent the epididymis) can be fluid-filled with secretions that are androgen-dependent and are thus an integral part of the measured endpoint. Prevention of leakage of these secretions in a consistent manner (by the use of a hemostat) as described in the guidelines is thus vitally important and a key means of reducing measurement errors and variability in weight of the retrieved tissues. It is also important that emphasis is placed on the importance of having the same individual do all of the dissections and weighings whenever possible. The same document tabulates acceptable variation in endpoint measurements (CVs); these are endpoints with naturally high variability and, compounded by variation in dissection/weighing, means that acceptable CVs are set in the 20-40% range for the most useable endpoint tissues. The most useful endpoint (SVCG) based on the evaluation studies is also one of the most variable. Data analysis is described in detail and for participating laboratories will undoubtedly require expert statistical input.

Dr. Talsness. The protocol is fairly detailed and provides significant information to perform the assay successfully. Information regarding excision of the sex accessory tissues is provided in the guidelines for the castrated version and a description for the testes and epididymides should be included here. Guidance should be included in terms of the number of doses required for test substances and the information that greater doses most likely have to be employed in the weanling mode than in the castrated version. Criteria for acceptable test performance need to be determined and/or included in the document (comment based on report identified above). Additional discussion regarding acceptable standard curves (weanling version has a flatter dose response curve for TP) with reference substances and

that this should be repeated at appropriate intervals to ensure proper test performance. Finally, more information regarding the interpretation of data is required, e.g., what constitutes a positive response when only some of the tissues exhibit changes that are statistically significant.

D. The intra- and interlaboratory reproducibility of the test method should be demonstrated.

Data should be available revealing the level of reproducibility and variability within and among laboratories over time. The degree to which biological variability affects the test method reproducibility should be addressed.

Dr. Chahoud. Since the weanling assay is designed as a screening test to identify if a substance is an androgen receptor agonist or antagonist, intra- and interlaboratory variability regarding dose-response relationship and biology is of low importance. The important question is whether all laboratories were able to identify the substances as androgens or antiandrogens. In this case, all laboratories were able to identify the properties of the substances tested and therefore intra- and interlaboratory reproducibility is reasonable.

Dr. Eldridge. The Report contained an extensive presentation and analysis of reproducibility and variability within laboratories and between laboratories (3 labs in most cases). Variability over time was not a goal of the present study. There was also some presentation and discussion of data contrasts between the traditional Hershberger method (adult castrate) and the present method (weanling intact), which could be a relevant issue (see remarks in response to Item 2). In addition, there was extensive discussion of the consequences of inter- and intra-laboratory variability that may likewise be a factor when considering the present method.

Dr. Patisaul. Inter-laboratory variability and reproducibility is a significant concern. Inter-laboratory variability is unquestionably higher for the weanling version of the Hershberger assay than for the castrated adult version. There was a statistically significant effect of laboratory for nearly every compound and every endpoint tested including the effect of TP. This is particularly worrisome because it is the positive control group. This is likely due to the substantially decreased sensitivity of the assay and the increased technical skill required to dissect and weigh each tissue correctly.

Many of the labs found "marginally insignificant" effects. The Draft OECD Guideline for the Testing of Chemicals (Revised October 20, 2008) states that the study should be repeated when (page 8, line 320), "at least two target tissues were marginally insignificant, i.e. p values between 0.05 and 0.10). Under this requirement a few of the labs (Bayer Crop, Bayer Health, BASF and possibly Korea) would have to repeat the linuron (LIN) test (based on the data presented in Table 6.8 of the Immature Male Rat Hershberger Model Validation Report). Increasing the group size would likely be required to generate enough statistical power to make meaningful conclusions and avoid "marginally insignificant" results. Either way (increasing group size or repeating the assay) would require the sacrifice of more animals than the castrate version, an animal welfare issue that is of greater concern than castration.

The vast majority of compounds screened for endocrine disruption would likely be weak (anti-) androgens (for example, DDE), therefore the reliability and reproducibility of the bioassay for these types of compounds is absolutely critical. The data presented in the report do not sufficiently demonstrate that the weanling version of the Hershberger bioassay can reliably detect the effects of weak (anti-)androgens.

Dr. Sharpe. The variability of the test is adequately described for participating laboratories and, as would be expected, shows a considerable range with some labs performing better than others in a reasonably consistent way, as is the norm for any assay run in different laboratories. Variability over time was not evaluated at this stage. The use of coded standard samples is an excellent way of objectively evaluating assay performance in a blinded fashion (Tables 6.2, 6.5 and 6.8), although the fact that chosen doses of the test compounds in the earliest comparison were those that induced quite large effects (30mg/kg for flutamide, 100mg/kg for Linuron, 160mg/kg for DDE) perhaps makes this evaluation less of a test of how it would operate in practice with weaker anti-androgens. However, the follow-up blinded assessment using two expected negative compounds (DNP, NP) and a lower dose of flutamide (3mg/kg) gives some measure of reassurance of the utility of the assay as none of the 6 participating laboratories found any significant effect of either of the negative test compounds and all 6 laboratories found significant effects of the flutamide on the weights of androgen-dependent target organs (Table 6.5). A much more convincing evaluation was the 6-laboratory evaluation of coded anti-androgenic chemicals (DDE at 16 and 160mg/kg, Linuron at 10 and 100mg/kg) with results in Table 6.8. This showed that virtually all of the laboratories demonstrated significant effects of the higher doses of both compounds although not convincingly for the lower doses; this evaluation was most robust for SVCG but effects/trends were found for all of the other key endpoint organs (VP, LABC, COW). Biological variability is a big issue and is dealt with in several places in the review, but particularly relevant are the CVs obtained in the intact weanling evaluation studies (Tables in chapters 4 & 5) which show, in general, that the assay works with reasonable performance in this regard, especially when set against the acceptable criteria for each endpoint. This aspect is also covered in response to question 2 below.

Dr. Talsness. *Intra-laboratory reproducibility.* Three laboratories (Syngenta, Korea, Canada) repeated the following combinations:

- Vehicle vs 1 mg/kg TP
- 1mg/kg TP vs. 1mg/kg TP + 3 mg/kg FLU
- 1mg/kg TP vs. 1mg/kg TP + 100 mg/kg LIN
- 1mg/kg TP vs 1mg/lg TP+160 mg/kg DDE.

The changes in weights of the mandatory accessory sex tissues obtained in the two experiments were roughly compared for each laboratory for each of these combinations. The changes in tissue weights achieved by each laboratory for the individual tissues were roughly similar over time.

Exceptions were:
Vehicle vs 1 mg/kg TP
VP for Syngenta

1mg/kg TP vs. 1mg/kg TP + 3 mg/kg FLU
VP and COW for Syngenta
VP and SVCG for Korea

1mg/kg TP vs. 1mg/kg TP + 100 mg/kg LIN
SVCG for Korea and Syngenta

1mg/kg TP vs 1mg/lg TP+160 mg/kg DDE
VP for Syngenta and Canada
SVCG for Syngenta, Korea and Canada

In general, the laboratories were able to reproduce their results between the two experiments.

Inter-laboratory reproducibility. There is significant influence of the laboratory on the all of the organ weights regardless of the compound tested. However, all laboratories were able to achieve the same results at the highest doses tested. Where standard curves are available, not all laboratories were able to detect statistically significant changes in organ weights at the same doses (lower) tested.

It is expected that the biological variability in the weanling model would be higher than in the castrated model due to the low level of physiological testosterone and presence of an intact hypothalamic-pituitary-gonadal axis.

E. Demonstration of the test method's performance should be based on the testing of reference chemicals representative of the types of substances for which the test method will be used.

A sufficient number of the reference chemicals should have been tested under code to exclude bias.

Dr. Chahoud. The chemicals selected were appropriate and the number of reference substances tested is sufficient. The number of substances tested under code is reasonable.

Dr. Eldridge. Several typical representative chemicals were tested (androgen, androgen receptor antagonist, etc.). However, the spectrum of modes of action that natural and synthetic substances may use to interact with the Hershberger parameters is unknown and potentially great. Phase 3 of the present project did involve testing of coded unknown chemicals at 6 different labs, as a way to minimize bias.

Dr. Patisaul. A sufficient number of the reference chemicals were tested under code to adequately exclude the risk of bias.

Dr. Sharpe. The assay review includes assessment in different laboratories (3 or 6, depending on the phase) and testing of a candidate weak androgenic environmental chemical (Trenbolone), plus testing of two candidate weak anti-androgenic environmental chemicals (DDE, Linuron) as well as a positive anti-androgen (flutamide). This is not a huge test of the assay (4-5 anti-androgens would have been better), but as the evaluation has also included a blinded test and included evaluation of negative controls (as mentioned above), and these evaluations worked well, I am convinced that the assay works with reasonable sensitivity and specificity.

Dr. Talsness. The following chemicals were coded and evaluated in Phase 3: TP (agonist), FLU, DDE, LIN (antagonists), and DNP and NP (negative chemicals).

A coded agonist was not tested per se as some laboratories did not perform a vehicle and statistical analyses between TP and vehicle was not performed. It is expected that more chemicals exhibit anti-androgenic effects than androgenic and priority was placed on the correct reference chemicals.

DNP and NP

None of the laboratories detected statistically significant changes in organ weights following exposure to the coded negative chemicals and all demonstrated significant changes with the positive control.

DDE
Androgen antagonism was not detected by any laboratory for the low dose of coded DDE.
Androgen antagonism was detected for all mandatory tissues by all laboratories for only the high dose of coded DDE.

Linuron
No laboratory detected androgen antagonism with the low dose of coded linuron.
Only one laboratory detected androgen antagonism with the high dose of coded linuron in the VP.
5/6 detected androgen antagonism with the high dose of coded linuron in the SVCG.
3/6 detected androgen antagonism with the high dose of coded linuron in the LABC.
1/6 detected androgen antagonism with the high dose of coded linuron in the COW.
On a per laboratory basis, 1 laboratory detected 4 statistically significant organ weight changes for the mandatory tissues indicative of antagonism

2 laboratories detected 2 statistically significant organ weight changes for the mandatory tissues indicative of antagonism

1 laboratories detected 1 statistically significant organ weight change for the mandatory tissues indicative of antagonism

2 laboratories did not detect any statistically significant organ weight changes for the mandatory tissues indicative of antagonism

Depending on the criteria employed (e.g., detection in at least 2 tissues), 50% of the laboratories did not identify linuron correctly.

The false positive rate of the two chemicals tested was 0 and the false negative rate for detection of weak anti-androgens may need improvement particularly as this test is to function as a screen where sensitivity is a priority over specificity.

In the phase 3 testing for the castrated model
(http://www.oecd.org/dataoecd/49/34/37479136.pdf),

6/10 laboratories detected 5 statistically significant organ weight changes for the mandatory tissues indicative of antagonism
3/10 laboratories detected 4 statistically significant organ weight changes for the mandatory tissues indicative of antagonism
1/10 laboratories detected 3 statistically significant organ weight changes for the mandatory tissues indicative of antagonism

Depending on the criteria employed (e.g., detection in at least 2 tissues), all laboratories were able to correctly identify linuron at the 100mg/kg/d dose indicating better sensitivity for the castrated version.

F. The performance of the test method should have been evaluated in relation to relevant information from the species of concern, and existing relevant toxicity testing data.
In the case of a substitute test method adequate data should be available to permit a reliable analysis of the performance and comparability of the proposed substitute test method with that of the test it is designed to replace.

Dr. Chahoud. The discussion of the data in light of the known effects of the test materials was sufficient for evaluation of performance of the Hershberger assay. Comparisons were presented between positive outcomes in the Hershberger screen and true adverse effects in developmental studies to indicate the appropriate species was used and the sensitivity of the methods appropriate.

The data allow the comparison between the weanling and the castrated version of the assay.

Dr. Eldridge. Chapter 7 was a rather brief but nevertheless very informative summary of comparisons between the present test (weanling intact) and the more typical, previously validated test (adult castrate). More will be said about those conclusions in discussion of Item 2.

Dr. Patisaul. The weanling version of the Hershberger bioassay is intended to replace the castrated adult version of the Hershberger bioassay. Chapter 7 of the Immature Male Rat Hershberger Model Validation Report summarized and compared data obtained using each version of the bioassay. Although there are many problems and caveats with this type of comparison (adequately stated in the report) it is an appropriate "first order" method for comparing the efficacy of both assays. CVs for the weanling version were included in the report, and CVs for the castrated adult version were obtained from the Draft OECD Guideline for the Testing of Chemicals (Revised October 20, 2008). Many of the endpoints in the weanling assay had CVs that were higher than the desired maximum. It does not appear that the two versions of the assay were compared in the same laboratory, at the same time, by the same individuals, which would be the best and most effective way of comparing the performance of each. Therefore it is difficult to determine if/how variables such as strain differences, timing of dose (morning or afternoon), inexperience, diet, or other potential confounds contributed to the outcome differences observed between the two versions.

The sensitivity of the two versions is clearly different, with the castrated version being significantly more sensitive than the weanling version. The weanling version requires a higher dose of TP and the degree to which organ weights change following exposure is clearly smaller. A smaller effect size likely means that larger group sizes are needed to obtain sufficient statistical power. This would mean sacrificing more animals, which defeats the purpose of adopting the weanling version over the castrated version for animal welfare purposes. CVs in the weanling version are generally acceptable but beyond maximum allowable limits in many cases, even within the control group. This likely stems from the technical skill required to properly dissect and weight the weanling organs, which is substantially smaller than the corresponding organs in the adult castrate.

Dr. Sharpe. The position and relevance of the test in relation to other tests of androgens/anti-androgens in rats is well and accurately described and is embedded in the toxicity literature. The fact that this (modified) assay is based on an assay that has been used by pharmaceutical companies for many years specifically for the evaluation of androgenicity and anti- androgenicity in compounds demonstrates that it has a rock-solid foundation. Its relative utility when compared with existing assays is covered in more detail in the response to Q2 below.

Dr. Talsness. Evaluation of the weanling version of the Hershberger Bioassy was conducted in a similar fashion (same doses and methodical procedure) to the castrated version allowing for comparison of the two methods. In general, sufficient data is available to allow comparison of the two methods. Analysis of a 5α-reductase inhibitor to assess detection of compounds with this mode of action would have allowed direct comparison of the three types of substances evaluated in the castrated version. Finally, presentation of GP

data from the castrated version would have been helpful of the reviewer to assess whether pertinent information was lost with loss of this endpoint.

G. Ideally, all data supporting the validity of a test method should have been obtained in accordance with the principles of GLP.

Aspects of data collection not performed according to GLP should be clearly identified and their potential impact on the validation status of the test method should be indicated.

Dr. Chahoud. GLP only provides information about the procedure applied in the experimental process. With respect to the scientific value of resulting data, the validity of tests performed according to GLP is not higher than studies not conducted under GLP conditions. The results were consistent between laboratories, and therefore the impact from the laboratories that did not conduct the analyses in accordance with GLP was not sufficient to raise concerns as to the validity of the validation effort.

Dr. Eldridge. Item 32 of Appendix B (OECD Model Protocols and Guidance) had the following statements:

> *32. Work should be conducted according to the principles of Good Laboratory Practice [OECD Good Laboratory Practice and Compliance Monitoring (12)]. In particular, data should have a full audit trail and be retained on file. Data will be collected in a manner that will allow independent peer review and written records maintained.*

A statement of certification that the principles of GLP were followed was not found in the Report, but there is no reason to believe that the principles were not followed. The data collection and animal husbandry appear adequate.

Dr. Patisaul. It appears that GLP was followed for all studies.

Dr. Sharpe. As far as I can tell from the report, all studies were conducted according to GLP.

Dr. Talsness. No comment.

H. All data supporting the assessment of the validity of the test method should be available for expert review.

The detailed test method protocol should be readily available and in the public domain. The data supporting the validity of the test method should be organised and easily accessible to allow for independent review(s), as appropriate. The test method description should be sufficiently detailed to permit an independent laboratory to follow the procedures and generate equivalent data. Benchmarks should be available by which an independent laboratory can itself assess its proper adherence to the protocol.

Dr. Chahoud. The presentation of the report is adequate to make an assessment and all documentation needed to adequately evaluate the performance of the Hershberger assay is publicly available on the OECD website.

Dr. Eldridge. It is unclear whether the test protocol used for the present report can be located in the public domain. Results and data from the present studies were clearly and completely presented in the Report. As stated above, the details of methods were

sufficiently presented to enable other laboratories to conduct the tests, should this modification be adopted.

Dr. Patisaul. To my knowledge the peer review is being conducted properly and the protocol for both versions of the Hershberger assay are in the public domain. A few details should be added to the protocol, as described in detail above in 1.C., including the time of day in which the compounds should be administered, the number of litters from which the weanlings should be collected, and the use of a phytoestrogen-free diet.

Dr. Sharpe. The assay review and supporting documents provide sufficient detail for all aspects of the assay, its operation, its methods and analysis to enable its introduction and use by a new laboratory. Benchmark data and statistics are provided that should enable a new laboratory to evaluate its own performance, and I presume that if the assay is approved and adopted then mechanisms will be established to allow for the testing of coded samples or samples for quality control by new laboratories.

Dr. Talsness. The data for this review was presented in a logical and organized fashion. The review could be conducted efficiently and information was easy to find. Suggestions for additional information for the test method have been described above.

2. Given that OECD is considering to apply the same performance criteria for both assays (the same upper CV limits, derived from the castrate version of the assay), is the weanling version of the Hershberger an adequate substitute for the traditional castrate version of the Hershberger assay?

Dr. Chahoud. Both versions of the Hershberger assay are performed as a screening test for the identification of androgenic or antiandrogenic properties of chemicals. Taking this into account the comparison between both versions should be based on the following considerations:

- *Can both versions identify the substances of interest?* Yes, identification is possible in both versions.

- *Is it necessary to identify dose-dependent effects?* No, since the purpose of this test is the screening of substances. Although, the castrated version is more sensitive, the weanling version can also be employed.

- *What is the advantage of the weanling version?* It is a clear advantage that in the weanling version a biologically intact system can be used and no interference with physiological processes is necessary. Furthermore, the procedure of castration has considerable impact on animal welfare and can be avoided in the weanling version, which is also less time and cost consuming.

Dr. Eldridge. The key word in this question is “adequate”.

The Report of the substitute assay (weanling intact male, or WI) identifies a number of attributes that contrast its performance with the traditional Hershberger assay (adult castrate male, or AC).

Advantages:

A. *The WI assay does perform.* Although the immature rat has already been exposed to some endogenous androgen, the accessory sex organs are still sufficiently rudimentary that they can measurably respond to administered androgen. Thus, the most basic question, i.e. can WI male rats be used to identify an androgenic substance administered in a screening assay is answered in the affirmative.

B. *No surgical procedure.* The WI method does not include the orchidectomy steps that can have attendant issues of anesthesia, post-surgical complications, infections, additional training and labor for technical staff, plus the need to consult institutional animal care and use guidelines more closely. The animal subjects are simply shipped in from a supplier, administered test substances according to a protocol, euthanized and necropsied. Because a Hershberger model is under consideration for testing of hundreds, even thousands, of substances, this smaller "footprint" of the WI model for animal rights concerns would be advantageous.

C. *Somewhat lower cost.* Rat purchase cost rises progressively with advancing age; weanling animals would be significantly less expensive than adults of 2 or 3 months old. Smaller animals can be housed in fewer cages and handled more easily. Finally, the absence of the surgical step (discussed above in B) would obviate a number of ancillary costs, not the least of which is labor paid for technical staff training and performance of the orchidectomies.

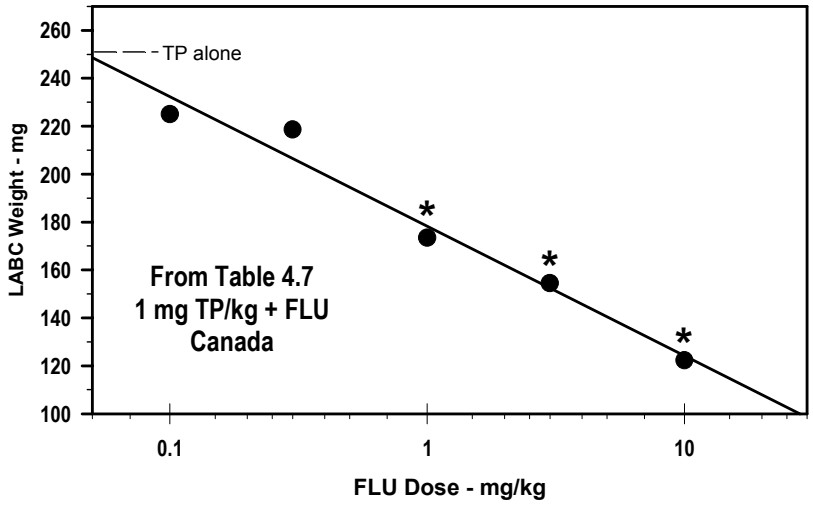
D. *Additional parameters to measure.* Examination of intact animals offers some measures not traditionally expected in a Hershberger model: testis weight and histology, and levels of testicular and pituitary hormones. Androgen agonists, antagonists and modifiers might be expected to influence some of these additional parameters. The WI would offer not only more targeted responses, it would also provide more clues for additional testing that would likely follow the identification of positive responses in the initial screen.

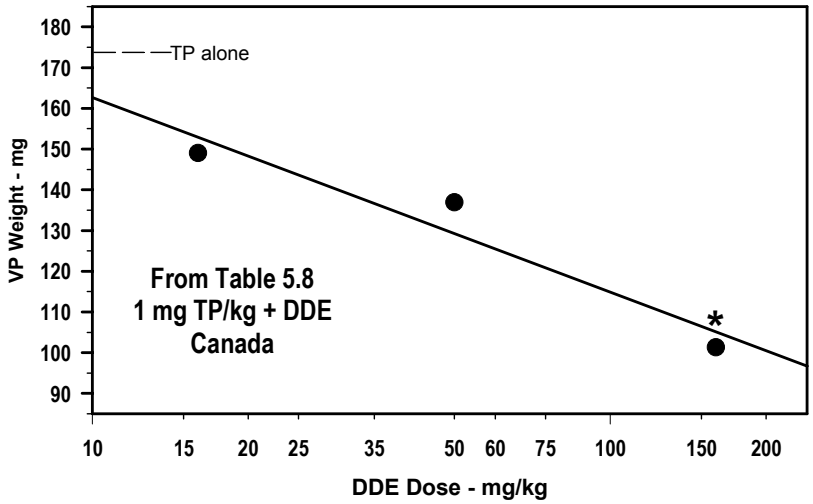
Disadvantages:

A. *Substantially lower signal-noise ratio.* Table 7.1 in the Report presents some critical data showing that the WI protocol is much less sensitive than the traditional AC: response of the ventral prostate weight to 0.4 mg testosterone propionate (TP) was 8.84 x baseline in AC animals but only 1.29 x baseline in WI animals. At a higher TP dose of 0.8 mg, AC was 11.86 x baseline while WI was only 1.53 x baseline. This is a serious loss of sensitivity, *to a known potent androgen.* A weaker androgen might be barely discernable, and will likely require much higher doses to be recognized, in the WI model compared to the AC model. While Advantage A was outlined above ("The WI assay does perform"), from a strictly biological standpoint, one might also suspect that the more sensitive AC assay would succeed in some instances when the WI would fail.

B. *Necropsy dissection is more critical.* The Report engages considerable discussion of coefficients of variation (CV) and parametric variability in general, both within labs and between labs. The typical group size for each administered substance is 6. Most of the measures in a Hershberger assay are organ weights that require very careful (precise) dissection and cleaning of tissues before weighing. If the WI model produces weight changes that are only 50-100% from baseline, group variances of 20-30% might render a true change statistically insignificant, that should be clearly identified in the AC model.

C. *Two examples of this problem are found in the figures below.* These are data plotted from two tests of androgen antagonism. These graphs show that, although the increasing antagonist doses produced a steady, linear decline of weight, the mean changes at the lower doses were not statistically significant. In both cases the lowest dose producing a significant change required a decline of greater than 30% from the mean a zero dose of antagonist. The trends clearly plotted effects at lesser





doses that remained undetectable due to group variances.

D. *Condition of prepubertal accessory organs.* While there is no question that target organs of WI animals responded similarly and predictably to those of AC animals, it remains possible that the rudimentary organs of young animals exposed to small amounts of endogenous hormones are not quite the same as atrophied organs of adult animals that were fully developed prior to orchidectomy. It remains unknown whether some organs might respond quite differently (WI versus AC) to non-classical agonists or antagonists. Although the AC model is by no means the ideal set of targets for evaluating responses to androgen agonists and antagonists, the model has been used for so long that it has been accepted as a "reference test".

E. *Acquisition of test subjects.* On page 15 (Section 3.0) of the Report is the following under the subheading "Age and Acclimitasition":

> ***Young weanling animals should be employed in a relatively small time window between weaning and before puberty (i.e., PND 20 to 34)…… The treatment with initiation of dosing (on study) may commence as early as pnd 21 days of age, but preferable not later than pnd 24. The laboratory is allowed some flexibility to schedule the experimental work efficiently.***

The ideal situation is to test a substance on a group of immature male rats born on exactly the same day, but it is not always possible for a supplier to provide 36 or 42 animals of *exactly* the same age. Shipments may frequently contain a few animals that vary from the requested age. Furthermore, labs will find it difficult to acquire such homogeneous groups on a repeated basic. The Report data support this premise, On Table 6.5 (page 100) are results from a study conducted at 6 different sites. In animals administered 1 mg/kg of TP, the group mean body weights range from 85.9 gm to 150.2 gm, and % CVs ranged between 4.5% and 13.5%. The study groups were decidedly not homogeneous. When groups of animals are acquired as adults, a few days' difference of age has much less effect on homogeneity. Animal ordering for the WI model will require more rigorous standards when dealing with suppliers, who may struggle to meet such exacting demands.

F. *Complications resulting from intact pituitary-testicular axis.* The traditional AC model of Hershberger has no testes present and open feedback loops governing pituitary hormone secretion. The presence of testes in the WI model could offer some unexpected or non-specific results. For example, an inhibitor of Leydig cell steroidogenesis could diminish endogenous testosterone secretion and appear to be a weak antagonist at the androgen receptor. This could lead to an erroneous positive result that would cloud results from other tests. An agent that stimulates gonadotropin secretion could enhance testosterone output and be identified as an androgen agonist. The fundamental Hershberger model was designed as an animal devoid of actual or potential endogenous androgens.

G. *Thyroid in immature versus adult males.* As mentioned in the Report summary, it is uncertain whether the thyroid axis of adult male rats functions in the same manner as that of the weanling, with respect to expression of gonadal and accessory organ development and function.

Dr. Patisaul. It does not appear that there is sufficient scientific or animal welfare justification for using the weanling version of the Hershberger bioassay in place of the castrated peri-pubertal/adult castrate version. The sensitivity of the weanling assay is demonstrably lower and the degree to which the organ weights change following (anti-)androgen administration is significantly smaller. The magnitude of each effect is also substantially smaller and, by extension, so is the margin for human error when collecting the organs and obtaining the weights. These limitations alone are sufficient justification for rejecting the suggestion that the weanling version is an adequate, appropriate, and effective substitute for the castrate version.

It appears that animal welfare concern over the castration procedure prompted the evaluation of the weanling version as a potential alternative. Unfortunately, the lower sensitivity of the weanling version compounded by the small effect size of each endpoint raises the likelihood that larger group sizes would be needed to overcome inter- and intra-laboratory variability and achieve reliable, accurate and repeatable results. Increasing

group sizes would require the sacrifice of more animals, a more substantive animal welfare concern than castration. Pain and discomfort associated with castration is easily managed and routinely done so for millions of laboratory animals, companion animals, and livestock.

The only situation where the weanling version may be advantageous over the castrate version is where an alternative mechanism of action (something other than direct interaction with ARs) is suspected. In this case, using a gonadally intact individual would be of interest because the HPG axis could respond to the exposure. However, the high dose of TP needed for the weanling version would likely suppress HPG function, thereby confounding this potential advantage. So even in this specific situation, although it would be better than the castrate version of the Hershberger assay, it would not likely be sensitive enough to sufficiently identify weakly (anti-) androgenic compounds.

In general, the adult castrate version of the Hershberger bioassay has proven to be a reliable, *in vivo* method for screening (anti-)androgenic compounds. Although the use of weanlings would eliminate the need for surgical castration, saving time for the technicians and avoiding this animal welfare concern, use of the weanling version introduces other complications and animal welfare concerns. These include increased difficulty in obtaining and weighing the organs correctly (because of their smaller size), reduced overall assay sensitivity (thus increasing the likelihood that more animals will need to be sacrificed), and increased risk of needing to repeat the assay or perform a different assay to confirm the findings. Therefore the weanling version of the Hershberger bioassay is not an adequate substitute for the traditional castrate version.

Dr. Sharpe. There are three considerations in answering this question: comparative specificity of the two assays, comparative sensitivity of the two assays and comparative variability of endpoints in the two assays (comparison of CVs for the same endpoints). Overall conclusions are then drawn.

Comparative specificity of the intact weanling assay versus the castrate rat assay. The intact weanling assay showed high specificity for androgens, anti-androgens and inactive compounds, although this is based on testing of a very limited number of compounds when compared with the extensive (historical) evaluation that has been undertaken using the castrated rat version of the assay. However, considering the near-identical basis for the two assays, there is no reason to expect that there will be any major difference in specificity of the intact weanling assay in comparison with the castrated rat assay.

Comparative sensitivity of the intact weanling assay versus the castrate rat assay. One of the main reservations expressed about the intact weanling Hershberger screening assay is its lower sensitivity to detect androgenic or anti-androgenic chemicals when compared to the castrated adult or pubertal rat versions of the Hershberger assay (discussed below). This is a pivotal point as the optimal screening assay is one that has the minimum false negative detection rate, and lack of sufficient assay sensitivity would be likely to increase the false negative rate. This raises the issue of what determines assay sensitivity? In the intact weanling assay it will be determined by the combined influences of:

1. The sensitivity of the target (androgen-dependent) organs to androgen action, which in turn will be determined by the number of androgen receptors (AR).

2. The magnitude of response of the target tissues (i.e. the weight range within which the target organ can be changed).

3. For anti-androgens, the level of stimulation by testosterone/other androgens, this representing the summation of exogenously administered testosterone propionate (TP) plus any endogenous androgens.

Item 1 is essentially fixed, although the available evidence suggests that androgen exposure will perhaps upregulate AR expression. Items 2 and 3 are to a large extent inter-related but are discussed separately for clarity.

Item 2 has a wide range for androgenic chemicals but the range for anti-androgenic chemicals is determined by the dose and duration of TP administration. Sufficient TP stimulation needs to have been applied to give a good working range in which anti-androgenic chemicals can effectively antagonize. This was a key goal of the validation/set-up studies in which a TP dose-response curve was undertaken. As with optimizing all assays, a key aim is to ensure that it is not operated at the upper end (or in excess of the maximally-stimulating dose) of the dose-response curve but rather on the most linear (steepest) part of the curve. In general, experience shows that operating towards the middle of the steepest part of the curve gives the greatest potential for 'unknowns' to push the response up or down steeply and with reasonable sensitivity. This is the situation for an assay that is designed to measure activity in unknowns, whereas in the intact weanling assay for anti-androgens, a change only in one direction (downwards) is being sought for, and therefore the minimum dose (1 mg/kg) of TP that gave close to a maximal stimulation of androgen target organs was used, so as to provide the maximum working range within which test anti-androgenic chemicals could exert their effect. Though this is reasonable thinking, it does to an extent ignore the issue of assay sensitivity. In essence, the more androgen that a test anti-androgenic chemical is competing with (for binding to a fixed number of AR) at androgen target organs, the smaller an effect it will have and thus the lower the sensitivity of the assay. In practice, in setting up an assay there is inevitably a 'trade off' between assay sensitivity and magnitude of effect because, for example, if the intact weanling assay was operated towards the lower end of the TP dose-response curve then the magnitude of effect that a test anti-androgenic chemical could have would be more limited, which would likely result in a higher CV and greater difficulty in detecting an effect. In many respects, assay sensitivity is usually tailored according to its purpose. For the OECD, two concerns have to be taken into account, namely assay sensitivity and robustness, bearing in mind that the assay needs to be easily useable in many different laboratories, some of which may lack prior experience.

It is clear from Table 4.5 that the 0.8mg TP dose gave nearly the same degree of increase in weight of target organs (and thus the working range for the assay) as did 1mg TP and it was stated towards the end of this section that both doses would be studied further when evaluating the assay using test compounds. In fact only data for 1mg TP is subsequently shown, so I presume that 0.8mg was not tested further. This decision is a little odd as, for the reasons detailed above, use of the 0.8mg TP dose would have been the more logical choice as it could only make the assay more sensitive compared with 1mg TP without affecting the working range of the assay or increasing CVs (Table 4.5); the increase in sensitivity would probably have been modest but I am surprised at why the 1mg dose was preferred for use.

With only minor exceptions and at every dose of flutamide tested, the castrate rat Hershberger assay showed a notably larger suppressive effect than did the weanling assay, especially for VP and SVCG (Table 7.2). The magnitude of difference was, however, most pronounced for the higher doses of flutamide (3 and 10mg/kg) and was more marginal for the lower doses (0.1 and 0.3 mg/kg). It is these lower doses that are more relevant to the effective screening of weak anti-androgenic environmental compounds. In general, when the sensitivity of the two assays to detect a weak androgenic (Trenbolone) and two weak

anti-androgenic (DDE, Linuron) chemicals was compared, the differences found largely reflected expectations based on the flutamide dose-response. Thus, for comparison of the two assay's abilities to detect Linuron (3, 10, 30 or 100mg/kg), there was only minor evidence for higher sensitivity of the castrate versus the intact weanling assay (Table 7.4), and only the top dose of Linuron (100mg/kg) had reasonably consistent significant effects in both assays. A similar comparison for DDE (5, 16, 50, 160mg/kg) provided clearer evidence of higher sensitivity of the castrate assay compared with the weanling assay, as the former detected anti-androgenic effects of 50mg/kg DDE in 1 or more/all of the participating laboratories for all 4 target organs (VP, SVCG, LABC, COWS) (Table 7.5). In contrast, in the weanling assay at this dose, a significant effect was only detected for the most robust endpoint (SVCG), and then only in 2 out of 3 laboratories; no significant effects were detected for the VP, LABC or COWS (Table 7.5). In both assays, significant effects were detected for all 4 organs at the next highest dose (160mg/kg) of DDE (Table 7.5). Despite the sensitivity difference between the two assays, the overall perspective was that either trends or effects of DDE were detected in the intact weanling assay at doses of DDE that caused effects or trends in the intact castrate assay, so the difference in performance between the two assays was a relatively modest one.

One potential reason for differing sensitivity of the two assays is that, in the castrate assay, the anti-androgens are competing against administered doses of 0.2 or 0.4 mg/kg TP whereas in the weanling assay they are competing against an administered dose of 1mg/kg TP (plus any endogenously produced androgens, though this is likely minimal). This is bound to make a difference, though it is obviously not the full explanation.

Comparative variability of the intact weanling assay versus the castrate rat assay. For this comparison, I used data for the castrate Hershberger assay published by Owens et al (2007) in *Environmental Health Perspectives* (115: 671-678). Overall, there are no major differences in CVs for the various endpoints in the two assays. Therefore, at least for this rather limited comparison, I do not see endpoint variation in the intact weanling assay as being a limiting factor in its operation and utility. Arguably the most important evaluation of the intact weanling assay performance was in the 6-laboratory evaluation of coded anti-androgens (DDE and LIN at two doses each). This showed that measurement variation outside of the pre-set maximum allowable CV occurred sporadically rather than consistently (Table 6.10). Thus allowable CVs were exceeded once for VP (DDE 16mg/kg), once for SVCG (DDE 160mg/kg), three times for LABC (one at DDE 16mg/kg, two at DDE 160mg/kg) and 4 times for COWS (one at DDE 16mg/kg, two at DDE 160mg/kg, one at LIN 10mg/kg). The important messages from this evaluation are (1) the sporadic nature of these excesses means that it is not directly assay-related (i.e. there is not something intrinsically wrong with how the assay works); (2) the excesses occurred as frequently with higher doses of anti-androgens, which caused clear anti-androgenic effects, as they did at lower doses where effects were marginal; (3) the target organ identified from the evaluation studies as providing the most sensitive and robust response to both androgens and anti-androgens in the intact weanling assay, namely the SVCG, was the best performing endpoint (along with VP) in terms of meeting acceptable CVs (Table 6.10). Based on comparison of CVs, there is no reason to reckon that the weanling assay will perform any differently than the castrated rat assay.

Dr. Talsness. CVs for the data presented for the ***weanling*** model
-TP: standard curve (Table 4.4)
-Tren (Table 5.4)

The coefficients of variation of the data from all the laboratories used to evaluate Tren and to generate the standard curve with TP were below the maximum allowable CV's indicating

adequate performance off the assay. According to this standard, the three laboratories were able to successfully complete this portion of the experiment.

The following indicates where the maximum CV values were surpassed.
-TP+FLU (Table 4.9)
The CVs were exceeded by one laboratory for the COW

-LIN (Table 5.6)
The CVs were exceeded by one laboratory for the VP and the LABC and by one laboratory for the COW.

-TP + DDE (Table 5.10)
The CVs were exceeded by 2/3 laboratories for the LABC.

-TP + FLU (Table 6.7)
The CVs were exceeded by 3/6 laboratories for the COW.

Regardless of substance tested, the CVs for the SVCG were never over the maximum allowed.

CVs for the data presented for the ***castrate*** model

Table 11 in the phase 3 report for the castrate model indicates that for evaluation of agonists, the CV was surpassed for
-4/10 laboratories for the VP
-4/10 laboratories for the SVCG
-2/10 laboratories for the COW

Table 15 in the phase 3 report for the castrate model indicated that none of the 10 laboratories exceeded the maximum CV for the antagonist data for any of the mandatory accessory sex tissues.

The CVs for the GP and LABC were never over the maximum allowed.

Even though tissue dissection may be more challenging in the weanling model, it appears to be technically feasible as in most cases the laboratories were able to achieve CVs under the maximum limit. A hundred percent success rate was also not achieved in the castrated version although the animals are bigger and the tissues have been theoretically been primed to react in a more pronounced fashion. In this respect, the castrated version does not provide an advantage over the weanling model.

Sensitivity. Comparison of the weanling and castrated versions of the assay in terms of changes in tissue weight in response to exposure to testosterone propionate (Table 7.1) indicates that the castrated version of the Hershberger bioassay is more sensitive. Statistically significant changes in tissue weights were observed by all laboratories in all tissues starting at 0.2 mg TP /kg/d for the castrated version and at 0.8 mg TP/kg/d for the weanling version. It is also apparent that the dose response curve for the weanling version is much flatter. The relative changes in tissue weights are much smaller and probably accounts for the increased difficulty in detecting statistically significant changes in weight.

Comparison of the weanling and castrated versions of the assay in terms of changes in tissue weight in response to exposure to testosterone propionate and flutamide (Table 7.2) indicates that the castrated version of the Hershberger bioassay is more sensitive. Statistically significant changes in tissue weights were observed by all laboratories in all

tissues starting at 1 mg FLU /kg/d for the castrated version and at 10 mg FLU/kg/d for the weanling version.

Comparison of the weanling and castrated versions of the assay in terms of changes in tissue weight in response to exposure to trenbolone (Table 7.3) indicates that the castrated version of the Hershberger bioassay is more sensitive. Statistically significant changes in tissue weights were observed by all laboratories in all tissues at the highest dose (40 mg TREN /kg/d for the castrated version and only 2/4 tissues were determined to be statistically significant by all laboratories at the highest dose of 40 mg TREN /kg/d for the weanling version.

Comparison of the weanling and castrated versions of the assay in terms of changes in tissue weight in response to exposure to testosterone propionate and linuron (Table 7.4) indicates that neither version of the Hershberger bioassay performed particularly well in detecting antagonistic action with this substance. At the highest dose tested, all of the laboratories were able to detect a statistically significant change in tissue weight for only one tissue in both versions of the bioassay. However, the comparison in Table 7.4 is somewhat misleading as it is more important to know how a lab would identify a substance as to how many labs observed a change in a specific tissue weight. Looking at the original data on a laboratory basis reveals that ¾ labs would correctly identify linuron and one laboratory had an indication in one organ using the castrate model as

1 lab detected changes in all five tissues
2 labs detected changes in four tissues
1 lab detected changes in one tissue.

In the weanling model:

2 labs detected changes in three out of four tissues
1 lab detected changes in two out of four tissues.

Looking at the data on a per laboratory basis indicates that both versions were able to identify linuron.

Comparison of the weanling and castrated versions of the assay in terms of changes in tissue weight in response to exposure to testosterone propionate and DDE (Table 7.5) indicates that all laboratories detected a statistically significant change in tissue weight for two of the tissues at the dose of 50 mg DDE/kg/d using the castrated version of the Hershberger bioassay while a statistically significant change was not found by all of the laboratories for any of the tissue weights at this dose. All laboratories observed statistically significant changes for all tissues at the highest dose regardless of method. Again, the castrated version appears to be more sensitive.

General Comments

Dr. Chahoud. In conclusion, the substitution of the traditional castrate version of the Hershberger assay with the weanling version is justified.

Dr. Eldridge. The weanling version of the Hershberger assay appears to be a barely adequate substitute for the traditional castrate version. The WI model has several points in its favor. Aside from the political considerations, the most important advantage of the WI model is that it would probably be somewhat easier, simpler and less expensive than the AC model. This is not insignificant because, if the Hershberger is adopted as one of a

battery of mandatory endocrine disruptor screens conducted around the world, there will likely be many commercial and industrial laboratories attempting to perform it. While it is noble for governments to engage a process of determining whether environmental contaminants can adversely impact endocrine function of humans and wildlife, there is established concern that accurate, reliable testing will be problematic when performance is upregulated on a wide scale. It is critically important that these screens be as simple and as unambiguous as possible.

On the other hand, while the WI model takes one step forward, it takes two steps back: it appears to be less sensitive and it introduces some unknown variables.

Loss of sensitivity is a significant problem because endocrine disruptors are typically much weaker on a molar basis than native estrogens and androgens. There are reports of estrogen antagonists having measured affinities for estrogen receptor binding at 1 million times higher concentration than estradiol itself. With an *in vivo* Hershberger assay, unknown chemicals may need to be administered at levels approaching the maximum tolerated dose (MTD), yet it remains possible that androgen receptor signaling would be very weak even at MTD levels. The assay's measures need to be optimally sensitive to pick up a weak signal before non-specific toxicity is reached.

Said another way, the WI model risks a yield of "false negative" results that the AC model would detect. Above all, a screen must not miss a true disruptor, even if it is extremely weak. A possible conflicted scenario could occur if a substance which responded at high concentration (or low affinity) in one or more *in vitro* screens (e.g., an androgen receptor binding assay or an AR-linked reporter construct) were to produce negative results with a poorly-sensitive *in vivo* Hershberger screen, thus leading to a confused, uncertain regulatory process.

The problem of the additional variables could also become serious. As stated in earlier remarks, the Hershberger assay model was purposely designed around a castrate animal in order to clear "background noise" of endogenous steroids and a functioning hypothalamic-pituitary-gonadal axis.

The Report establishes that this noise is already a factor by diminishing the WI model sensitivity, which is further compounded by problems of making precise measures from dissection. In addition, the presence of testes and intact H-P-G feedback loops (which are very functional in the weanling male) retains some variables that may potentially confound analysis of results. "False positive" results could appear from an agent exerting non-endocrine mechanisms through the testis or in neuroendocrine organs. While possibly less significant than a false negative result, a false positive could send investigators on long and expensive second-level searches for mechanisms of hormonal effects that were erroneously suggested by the screen.

In summary, the weanling intact Hershberger model does not seem to be an improved substitute for the adult castrate model, nor does it appear to be even an equivalent substitute. It is a substitute that works, but which comes with a compromise of scientific quality. It is not completely inadequate; it remains barely adequate.

Dr. Patisaul. The Hershberger is a reliable, well validated assay for the screening of androgenic and anti-androgenic compounds. Traditionally, it uses castrated peri-pubertal or adult male rats. Castration is essential to ensure that endogenous androgen levels, which could significantly interfere with the outcome of the assay, remain as low as possible. Use of a gonadally intact weanling male, as described in the Immature Male Rat Hershberger Validation Report (dated November 30, 2008), has a number of drawbacks. The report

clearly and adequately summarizes the potential limitations of using male weanlings instead of adult castrates including reduced sensitivity of the assay, and increased risk of inter- and intra-laboratory variability because the tissues are smaller and more difficult to dissect.

Most notable, however, is that to achieve sufficient statistical power, the weanling version of the Hershberger bioassay would likely require larger group sizes than the castrate version. The Draft OECD Guideline for the Testing of Chemicals (Revised on October 20, 2008) states that (page 6, line 196), "castration reduces the numbers of animals required to screen for these endocrine activities." The principle rationale for using the weanling version of the Hershberger assay instead of the castrated adult version is the desire to avoid castration for animal welfare purposes. Castration is a simple surgery that is performed routinely on laboratory animals as well as companion and livestock animals with minimal risk of complication or long term pain. It is my opinion that switching to a weanling model, which may require the sacrifice of more than twice as many animals, would be a vastly more significant animal welfare concern. There are no substantive scientific or animal welfare benefits of using the weanling version of the Hershberger Assay instead of the castrated adult version, but numerous clear drawbacks including lower sensitivity of the assay and the need to sacrifice significantly more animals.

Dr. Sharpe. The intact weanling assay is fit for purpose and can provide an adequate substitute for the castrated rat assay. The latter is slightly more sensitive and provides a wider working range of target organ weights, but its overall operation is not markedly different from that of the intact weanling assay. Arguably, when large numbers of compounds are screened, it is likely that the intact weanling assay may not detect some weak anti-androgens that the castrate adult assay does detect, but this is likely to be infrequent. This possibility is minimized by the use of multiple evaluation endpoints, and this is an inherent strength of the intact weanling assay, just as it is for the castrated rat assay.

It is stated in the evaluation report that one advantage of the intact weanling assay is that having an intact hypothalamic-pituitary-testicular (HPT) axis, means that it may detect a wider range of anti-androgenic compounds than the castrated rat assay. This remains only a theoretical possibility as there are no demonstrations that this is indeed the case – in my opinion, the HPT axis will be essentially shut down due to the TP treatment. One present drawback with the intact weanling assay is that there is very limited history in its use, whereas there is a huge depth of experience with the castrated rat assay.

Dr. Talsness. In summary, there are examples in both bioassays where the maximum coefficients of variation were exceeded. In this respect, the weanling version is an adequate substitute for the castrated version. Interestingly, CVs were exceeded with testing of antagonists, but not agonists in the weanling version and the opposite is true for the castrated version.

Since it is expected that there are more environmental antagonists than agonists, this may represent a possible advantage for using the castrated version of the bioassay. Based on the limited data, a preliminary deliberation is that it may be prudent to using the castrated version of the bioassay for the cases where the *in vitro* data indicate antagonist androgen action and reserve the weanling when agonist activity is expected.

Another advantage of the castrated version of the bioassay is that significant changes in tissue weights were detected at lower doses of the test compounds than in the weanling version. The increased sensitivity is probably due to the greater relative changes in organ

weight. Higher doses would have to be employed in the weanling version of the bioassay which may increase the number of false positives due to non specific action.

The advantages of the weanling version include an intact biological system and avoidance of the surgical castration procedure.

Although the castrated version is more sensitive, there is no example in the data provided where the compound was not eventually identified in the weanling version as in the castrated version and, therefore, the weanling version could be used as a substitute for the castrated version with the caveat that higher doses may have to be employed. The choice of which version to employ could be based on available *in vitro* data.

Appendix A. Original Submissions from Peer Reviewers

Dr. Ibrahim Chahoud [as-received]

INDEPENDENT PEER REVIEW OF THE WEANLING VERSION OF THE HERSHBERGER ASSAY

Ibrahim Chahoud

The following comments are based solely on the Validation Report dated November 30, 2008.

1. Comment on the adequacy of the validation programme by stating how well it meets the following validation criteria:

A. *The rationale for the test method should be available.*
This should include a clear statement of the scientific basis, regulatory purpose and need for the test.

Scientific basis: The scientific basis for the Hershberger Assay is well documented and employed for decades since the 1930's. The aim of the test is the evaluation of the ability of chemicals to change the weight of androgen-dependent tissues in prepubertal male rats. The test period lasts from approximately PND 21-33. At this age the tissues evaluated in male rats exhibit androgen receptors and at the same time serum testosterone concentrations are low. Therefore, the androgen dependent tissues are sensitive to exogenous androgens as well as anti-androgens. In contrast to the castrated version, the advantage of this model is the intact hypothalamic-pituitary-gonadal axis which provides a physiologically normal test situation.

Regulatory purpose: The regulatory need exists to rapidly assess and evaluate a chemical as a possible androgen agonist or antagonist or 5α-reductase inhibitor. The Hershberger bioassay serves as a mechanistic *in vivo* screening assay and its application should be seen in the context of the "OECD Conceptual Framework for the Testing and Assessment of Endocrine Disrupting Chemicals" (annex 2). The assay is part of Level 3 and is designed to provide data about a single endocrine mechanism, i.e. (anti)androgenicity. The inclusion in a battery of *in vitro* and *in vivo* tests to identify substances with potential to interact with the endocrine system is recommended, ultimately leading to hazard and risk assessments for human health or the environment.

Need for the test: Whether a substance acts as androgen or antiandrogen can be assessed in an *in vitro* test battery. However, the disadvantage of *in vitro* tests is their lack of information about the toxicokinetic properties of a given substance. Therefore, there is a need for *in vivo* tests in order to be able to get important information about resorption, metabolisation and elimination of the tested compound. The Hershberger Assay is an adequate test system for these purposes. Evaluation of the test as an alternative protocol to the castrated model version is justified.

B. *The relationship between the test method's endpoint(s) and the (biological) phenomenon of interest should be described***.**

This should include a reference to scientific relevance of the effect(s) measured by the test method in terms of their mechanistic (biological) or empirical (correlative) relationship to the specific type of effect/toxicity of interest. Although the relationship may be mechanistic or correlative, test methods with biological relevance to the effect/toxicity being evaluated are preferred.

The aim of the weanling version of the Hershberger Assay is the identification of androgen receptor agonists/antagonist and 5α-reductase inhibitors. The endpoints required are weight changes in androgen-dependent tissues (Cowper's glands, levator ani-bulbocavernosus muscle complex, seminal vesicles with coagulating glands and their fluids and ventral prostate). Optionally, other reproductive organs (testes, epididymides) as well as liver, kidney, adrenal glands and hormone concentrations are recommended.

Weight changes in androgen-dependent reproductive tissues are sensitive and relevant markers for androgen agonists and antagonists. Agonists will cause an increase in the weights of the respective tissues, while treatment with antagonists, will prevent the androgen-dependent organ weight increase. The choice of rat accessory sex organs is biologically relevant, mechanistically sensitive and adequate to detect the effects of androgen receptor agonists, antagonists and 5 α-reductase inhibitors.

The Hershberger assay is currently the best available *in vivo* assay for detecting androgen receptor agonists and antagonists. However, it should be stressed that this bioassay can only be used as a screening test.

C. A detailed protocol for the test method should be available.

The protocol should be sufficiently detailed and should include, *e.g.,* a description of the materials needed, such as specific cell types or construct or animal species that could be used for the test (if applicable), a description of what is measured and how it is measured, a description of how data will be analyzed, decision criteria for evaluation of data and what are the criteria for acceptable test performance.

The protocol is clearly written, comprehensive and there are adequate descriptions of the materials, equipment and methodology. Furthermore, performance criteria and justification of animal numbers are acceptable. The description of data analysis is comprehensive in the use of correct statistical approaches.

It is not clear if a statistically significant reduction in only one of the tissues weighed would be sufficient to be called a positive signal or outcome?

D. The intra- and interlaboratory reproducibility of the test method should be demonstrated.

Data should be available revealing the level of reproducibility and variability within and among laboratories over time. The degree to which biological variability affects the test method reproducibility should be addressed.

Since the weanling assay is designed as a screening test to identify if a substance is an androgen receptor agonist or antagonist, intra- and interlaboratory variability regarding dose-response relationship and biology is of low importance. The important question is whether all laboratories were able to identify the substances as androgens or antiandrogens. In this case, all laboratories were able to identify the properties of the substances tested and therefore intra- and interlaboratory reproducibility is reasonable.

E. Demonstration of the test method's performance should be based on the testing of reference chemicals representative of the types of substances for which the test method will be used. A sufficient number of the reference chemicals should have been tested under code to exclude bias.

The chemicals selected were appropriate and the number of reference substances tested is sufficient. The number of substances tested under code is reasonable.

F. The performance of the test method should have been evaluated in relation to relevant information from the species of concern, and existing relevant toxicity testing data.
In the case of a substitute test method adequate data should be available to permit a reliable analysis of the performance and comparability of the proposed substitute test method with that of the test it is designed to replace.

The discussion of the data in light of the known effects of the test materials was sufficient for evaluation of performance of the Hershberger assay. Comparisons were presented between positive outcomes in the Hershberger screen and true adverse effects in developmental studies to indicate the appropriate species was used and the sensitivity of the methods appropriate.

The data allow the comparison between the weanling and the castrated version of the assay.

G. Ideally, all data supporting the validity of a test method should have been obtained in accordance with the principles of GLP.
Aspects of data collection not performed according to GLP should be clearly identified and their potential impact on the validation status of the test method should be indicated.

GLP only provides information about the procedure applied in the experimental process. With respect to the scientific value of resulting data, the validity of tests performed according to GLP is not higher than studies not conducted under GLP conditions. The results were consistent between laboratories, and therefore the impact from the laboratories that did not conduct the analyses in accordance with GLP was not sufficient to raise concerns as to the validity of the validation effort.

H. All data supporting the assessment of the validity of the test method should be available for expert review.
The detailed test method protocol should be readily available and in the public domain. The data supporting the validity of the test method should be organised and easily accessible to allow for independent review(s), as appropriate. The test method description should be sufficiently detailed to permit an independent laboratory to follow the procedures and generate equivalent data. Benchmarks should be available by which an independent laboratory can itself assess its proper adherence to the protocol.

The presentation of the report is adequate to make an assessment and all documentation needed to adequately evaluate the performance of the Hershberger assay is publicly available on the OECD website.

2. Given that OECD is considering to apply the same performance criteria for both assays (the same upper CV limits, derived from the castrate version of the assay), is the weanling version of the Hershberger an adequate substitute for the traditional castrate version of the Hershberger assay?

Both version of the Hershberger assay are performed as a screening test for the identification of androgenic or antiandrogenic properties of chemicals. Taking this into account the comparison between both versions should be based on the following considerations:

- Can both versions identify the substances of interest? Yes, identification is possible in both versions.

- Is it necessary to identify dose-dependent effects? No, since the purpose of this test is the screening of substances. Although, the castrated version is more sensitive, the weanling version can also be employed.

- What is the advantage of the weanling version? It is a clear advantage that in the weanling version a biologically intact system can be used and no interference with physiological processes is necessary. Furthermore, the procedure of castration has considerable impact on animal welfare and can be avoided in the weanling version, which is also less time and cost consuming.

In conclusion, the substitution of the traditional castrate version of the Hershberger assay with the weanling version is justified.

Dr. J. Charles Eldridge [as-received]

RESPONSE TO PEER REVIEW CHARGES

for

INDEPENDENT PEER REVIEW OF THE WEANLING VERSION OF THE HERSHBERGER ASSAY

J. Charles Eldridge, Ph.D.
Wake Forest University

1. Comment on the adequacy of the validation program by stating how well it meets the following validation criteria:

A. The rationale for the test method should be available.

All of the test methods and the scientific basis for the methods were clearly presented. The needs for this test (Hershberger), the justification underlying the need to investigate the present modifications (use of intact weanling rats), and the regulatory purposes were also clearly presented.

B. The relationship between the test method's endpoint(s) and the (biological) phenomenon of interest should be described.

The test methods and the chosen parameters appear to be scientifically valid and biologically relevant. The report included a discussion of published literature and contained a good list of citations, both research papers and regulatory reports of assay validation. The targeted tests also appear to be relevant to both human health and to environmental safety of many species of male animals.

C. A detailed protocol for the test method should be available.

The methods were presented in good detail. It would not be difficult for a professional laboratory to conduct these tests in the manner intended by the methodology designers. Criteria for data evaluation and test performance were also described in detail, and appear to be reasonable.

D. The intra- and interlaboratory reproducibility of the test method should be demonstrated.

The Report contained an extensive presentation and analysis of reproducibility and variability within laboratories and between laboratories (3 labs in most cases). Variability over time was not a goal of the present study. There was also some presentation and discussion of data contrasts between the traditional Hershberger method (adult castrate) and the present method (weanling intact), which could be a relevant issue (see remarks in response to Item 2). In addition, there was extensive discussion of the consequences of inter- and intra-laboratory variability that may likewise be a factor when considering the present method.

E. Demonstration of the test method's performance should be based on the testing of reference chemicals representative of the types of substances for which the test method will be used.

Several typical representative chemicals were tested (androgen, androgen receptor antagonist, etc.). However, the spectrum of modes of action that natural and synthetic substances may use to interact with the Hershberger parameters is unknown and potentially great. Phase 3 of the present project did involve testing of coded unknown chemicals at 6 different labs, as a way to minimize bias.

F. The performance of the test method should have been evaluated in relation to relevant information from the species of concern, and existing relevant toxicity testing data.

Chapter 7 was a rather brief but nevertheless very informative summary of comparisons between the present test (weanling intact) and the more typical, previously validated test (adult castrate). More will be said about those conclusions in discussion of Item 2.

G. Ideally, all data supporting the validity of a test method should have been obtained in accordance with the principles of GLP.

Item 32 of Appendix B (OECD Model Protocols and Guidance) had the following statements:

32. Work should be conducted according to the principles of Good Laboratory Practice (OECD Good Laboratory Practice and Compliance Monitoring (12). In particular, data should have a full audit trail and be retained on file. Data will be collected in a manner that will allow independent peer review and written records maintained.

A statement of certification that the principles of GLP were followed was not found in the Report, but there is no reason to believe that the principles were not followed. The data collection and animal husbandry appear adequate.

H. All data supporting the assessment of the validity of the test method should be available for expert review.

It is unclear whether the test protocol used for the present report can be located in the public domain. Results and data from the present studies were clearly and completely presented in the Report. As stated above, the details of methods were sufficiently presented to enable other laboratories to conduct the tests, should this modification be adopted.

2. Given that OECD is considering to apply the same performance criteria for both assays (the same upper CV limits, derived from the castrate version of the assay), is the weanling version of the Hershberger an adequate substitute for the traditional castrate version of the Hershberger assay?

The key word in this question is "adequate".

The Report of the substitute assay (weanling intact male, or WI) identifies a number of attributes that contrast its performance with the traditional Hershberger assay (adult castrate male, or AC).

Advantages:

A. The WI assay does perform. Although the immature rat has already been exposed to some endogenous androgen, the accessory sex organs are still sufficiently rudimentary that they can measurably respond to administered androgen. Thus, the most basic question, i.e. can WI male rats be used to identify an androgenic substance administered in a screening assay is answered in the affirmative.

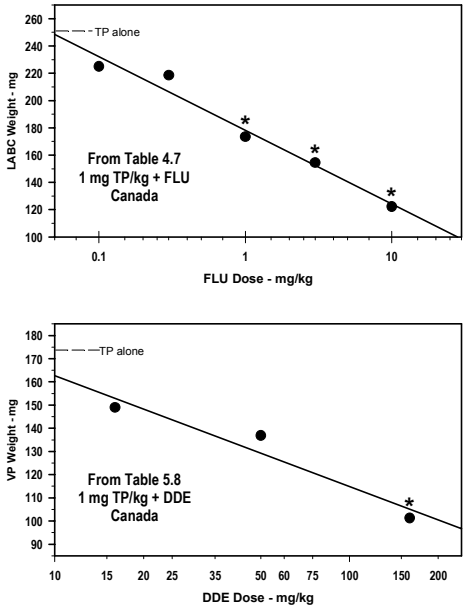
B. No surgical procedure. The WI method does not include the orchidectomy steps that can have attendant issues of anesthesia, post-surgical complications, infections, additional training and labor for technical staff, plus the need to consult institutional animal care and use guidelines more closely. The animal subjects are simply shipped in from a supplier, administered test substances according to a protocol, euthanized and necropsied. Because a Hershberger model is under consideration for testing of hundreds, even thousands, of substances, this smaller "footprint" of the WI model for animal rights concerns would be advantageous.

C. Somewhat lower cost. Rat purchase cost rises progressively with advancing age; weanling animals would be significantly less expensive than adults of 2 or 3 months old. Smaller animals can be housed in fewer cages and handled more easily. Finally, the absence of the surgical step (discussed above in B) would obviate a number of ancillary costs, not the least of which is labor paid for technical staff training and performance of the orchidectomies.

D. Additional parameters to measure. Examination of intact animals offers some measures not traditionally expected in a Hershberger model: testis weight and histology, and levels of testicular and pituitary hormones. Androgen agonists, antagonists and modifiers might be expected to influence some of these additional parameters. The WI would offer not only more targeted responses, it would also provide more clues for additional testing that would likely follow the identification of positive responses in the initial screen.

Disadvantages:

A. Substantially lower signal-noise ratio. Table 7.1 in the Report presents some critical data showing that the WI protocol is much less sensitive than the traditional AC: response of the ventral prostate weight to 0.4 mg testosterone propionate (TP) was 8.84 x baseline in AC animals but only 1.29 x baseline in WI animals. At a higher TP dose of 0.8 mg, AC was 11.86 x baseline while WI was only 1.53 x baseline. This is a serious loss of sensitivity, *to a known potent androgen*. A weaker androgen might be barely discernable, and will likely require much higher doses to be recognized, in the WI model compared to the AC model. While Advantage A was outlined above ("The WI assay does perform"), from a strictly biological standpoint, one might also suspect that the more sensitive AC assay would succeed in some instances when the WI would fail.



B. Necropsy dissection is more critical. The Report engages considerable discussion of coefficients of variation (CV) and parametric variability in general, both within labs and between labs. The typical group size for each administered substance is 6. Most of the measures in a Hershberger assay are organ weights that require very careful (precise) dissection and cleaning of tissues before weighing. If the WI model produces weight changes that are only 50-100% from baseline, group variances of 20-30% might render a true change statistically insignificant that should be clearly identified in the AC model.

C. Two examples of this problem are found in the figure to the left. These are data plotted from two tests of androgen antagonism. These graphs show that, although the increasing antagonist doses produced a steady, linear decline of weight, the mean changes at the lower doses were not statistically significant. In both cases the lowest dose producing a significant change required a decline of greater than 30% from the mean a zero dose of antagonist. The trends clearly plotted effects at lesser doses that remained undetectable due to group variances.

D. Condition of prepubertal accessory organs. While there is no question that target organs of WI animals responded similarly and predictably to those of AC animals, it remains possible that the rudimentary organs of young animals exposed to small amounts of endogenous hormones are not quite the same as atrophied organs of adult animals that were fully developed prior to orchidectomy. It remains unknown whether some organs might respond quite differently (WI versus AC) to non-classical agonists or antagonists. Although the AC model is by no means the ideal set of targets for evaluating responses to androgen agonists and antagonists, the model has been used for so long that it has been accepted as a "reference test".

E. Acquisition of test subjects. On page 15 (Section 3.0) of the Report is the following under the subheading "Age and Acclimitasition":

Young weanling animals should be employed in a relatively small time window between weaning and before puberty (i.e., PND 20 to 34)…… The treatment with initiation of dosing (on study) may commence as early as pnd 21 days of age, but preferable not later than pnd 24. The laboratory is allowed some flexibility to schedule the experimental work efficiently.

The ideal situation is to test a substance on a group of immature male rats born on exactly the same day, but it is not always possible for a supplier to provide 36 or 42 animals of *exactly* the same age. Shipments may frequently contain a few animals that vary from the requested age. Furthermore, labs will find it difficult to acquire such homogeneous groups on a repeated basic. The Report data support this premise, On Table 6.5 (page 100) are results from a study conducted at 6 different sites. In animals administered 1 mg/kg of TP, the group mean body weights range from 85.9 gm to 150.2 gm, and % CVs ranged between 4.5% and 13.5%. The study groups were decidedly not homogeneous. When groups of animals are acquired as adults, a few days' difference of age has much less effect on homogeneity. Animal ordering for the WI model will require more rigorous standards when dealing with suppliers, who may struggle to meet such exacting demands.

F. Complications resulting from intact pituitary-testicular axis. The traditional AC model of Hershberger has no testes present and open feedback loops governing pituitary hormone secretion. The presence of testes in the WI model could offer some unexpected or non-specific results. For example, an inhibitor of Leydig cell steroidogenesis could diminish endogenous testosterone secretion and appear to be a weak antagonist at the androgen receptor. This could lead to an erroneous positive result that would cloud results from other tests. An agent that stimulates gonadotropin secretion could enhance testosterone output and be identified as an androgen agonist. The fundamental Hershberger model was designed as an animal devoid of actual or potential endogenous androgens.

G. Thyroid in immature versus adult males. As mentioned in the Report summary, it is uncertain whether the thyroid axis of adult male rats functions in the same manner as that of the weanling, with respect to expression of gonadal and accessory organ development and function.

Summary

The weanling version of the Hershberger assay appears to be a barely adequate substitute for the traditional castrate version. The WI model has several points in its favor. Aside from the political considerations, the most important advantage of the WI model is that it would probably be somewhat easier, simpler and less expensive than the AC model. This is not insignificant because, if the Hershberger is adopted as one of a battery of mandatory endocrine disruptor screens conducted around the world, there will likely be many commercial and industrial laboratories attempting to perform it. While it is noble for governments to engage a process of determining whether environmental contaminants can adversely impact endocrine function of humans and wildlife, there is established concern that accurate, reliable testing will be problematic when performance is upregulated on a wide scale. It is critically important that these screens be as simple and as unambiguous as possible.

On the other hand, while the WI model takes one step forward, it takes two steps back: it appears to be less sensitive and it introduces some unknown variables.

Loss of sensitivity is a significant problem because endocrine disruptors are typically much weaker on a molar basis than native estrogens and androgens. There are reports of estrogen antagonists having measured affinities for estrogen receptor binding at 1 million times higher concentration than estradiol itself. With an *in vivo* Hershberger assay, unknown chemicals may need to be administered at levels approaching the maximum tolerated dose (MTD), yet it remains possible that androgen receptor signaling would be very weak even at MTD levels. The assay's measures need to be optimally sensitive to pick up a weak signal before non-specific toxicity is reached.

Said another way, the WI model risks a yield of "false negative" results that the AC model would detect. Above all, a screen must not miss a true disruptor, even if it is extremely weak. A possible conflicted scenario could occur if a substance which responded at high concentration (or low affinity) in one or more *in vitro* screens (e.g., an androgen receptor binding assay or an AR-linked reporter construct) were to produce negative results with a poorly-sensitive *in vivo* Hershberger screen, thus leading to a confused, uncertain regulatory process.

The problem of the additional variables could also become serious. As stated in earlier remarks, the Hershberger assay model was purposely designed around a castrate animal in order to clear "background noise" of endogenous steroids and a functioning hypothalamic-pituitary-gonadal axis.

The Report establishes that this noise is already a factor by diminishing the WI model sensitivity, which is further compounded by problems of making precise measures from dissection. In addition, the presence of testes and intact H-P-G feedback loops (which are very functional in the weanling male) retains some variables that may potentially confound analysis of results. "False positive" results could appear from an agent exerting non-endocrine mechanisms through the testis or in neuroendocrine organs. While possibly less significant than a false negative result, a false positive could send investigators on long and expensive second-level searches for mechanisms of hormonal effects that were erroneously suggested by the screen.

In summary, the weanling intact Hershberger model does not seem to be an improved substitute for the adult castrate model, nor does it appear to be even an equivalent substitute. It is a substitute that works, but which comes with a compromise of scientific quality. It is not completely inadequate; it remains barely adequate.

Dr. Heather B. Patisaul [as-received]

Critique Summary

The Hershberger is a reliable, well validated assay for the screening of androgenic and anti-androgenic compounds. Traditionally, it uses castrated peri-pubertal or adult male rats. Castration is essential to ensure that endogenous androgen levels, which could significantly interfere with the outcome of the assay, remain as low as possible. Use of a gonadally intact weanling male, as described in the Immature Male Rat Hershberger Validation Report (dated November 30, 2008), has a number of drawbacks. The report clearly and adequately summarizes the potential limitations of using male weanlings instead of adult castrates including reduced sensitivity of the assay, and increased risk of inter- and intra-laboratory variability because the tissues are smaller and more difficult to dissect.

Most notable, however, is that to achieve sufficient statistical power, the weanling version of the Hershberger bioassay would likely require larger group sizes than the castrate version. The Draft OECD Guideline for the Testing of Chemicals (Revised on October 20, 2008) states that (page 6, line 196), "castration reduces the numbers of animals required to screen for these endocrine activities." The principle rationale for using the weanling version of the Hershberger assay instead of the castrated adult version is the desire to avoid castration for animal welfare purposes. Castration is a simple surgery that is performed routinely on laboratory animals as well as companion and livestock animals with minimal risk of complication or long term pain. It is my opinion that switching to a weanling model, which may require the sacrifice of more than twice as many animals, would be a vastly more significant animal welfare concern. There are no substantive scientific or animal welfare benefits of using the weanling version of the Hershberger Assay instead of the castrated adult version, but numerous clear drawbacks including lower sensitivity of the assay and the need to sacrifice significantly more animals.

1.A. Rationale: The Immature Male Rat Hershberger Model Validation Report (dated November 30, 2008) clearly states that the "overall aim of the validation program is to demonstrate that the Hershberger bioassay is a robust, sensitive, reliable and reproducible bioassay that can be considered as the basis for an OECD Test Guideline. Once available, the test guideline is intended to be used as one element in an overall testing strategy for the detection and assessment of potential endocrine disruptors." Use of the Hershberger bioassay as a component of an endocrine disruptor screening program is appropriate and necessary.

The primary rationale for recommending the weanling version of the Hershberger Assay, instead of the castrate version, is the desire to avoid castration for animal welfare reasons. Another rationale for the switch is stated in the final paragraph of the report. Because the weanlings are gonadally intact, and therefore the hypothalamic-pituitary-gonadal (HPG) axis is theoretically capable of responding to the administration of an endocrine active compound, the weanling version of the Hershberger bioassay is potentially advantageous over the castrate version because it has the potential to detect effects resulting from disruption within the (HPG) axis in addition to direct action on androgen receptors. However, the high level of TP needed for the weanling version (1 mg/kg compared to 0.4 for the castrate version) would likely suppresses the responsivity of the HPG axis through steroid negative feedback. Therefore the ability of the HPG axis to adequately respond to the chemical insult under the test conditions of the Hershberger bioassay is likely minimal.

Another stated rationale for using the weanling version of the Hershberger bioassay over the castrate version is concern that the results obtained using the castrate version might not be extrapolable to gonad-intact animals. It is unlikely that either version would

produce results that are unequivocally extrapolable to the uncastrated adult (or juvenile), but this should not be considered a major problem because the primary goal of the assay is to screen for compounds with endocrine disrupting properties, not to make predictions about how the compound might affect gonadally intact individuals. For the purposes of screening, the assay with the greatest sensitivity, higher demonstrated degree of reproducibility, and lowest laboratory intervariability should be considered superior.

1.B. Endpoint: In general, the Immature Male Rat Hershberger Model Validation Report (dated November 30, 2008) clearly and adequately addresses the limitations of the weanling model, compared to the castrate model, the most significant of which is the potential for the (HPG) axis to respond to the compounds being screened and therefore interfere with the outcome. The most problematic drawback of using gonadally intact weanlings is that the sensitivity of the bioassay is clearly and markedly reduced. To compensate for this, greater numbers of animals would likely be needed to achieve sufficient statistical power which, to me, is a far greater animal welfare concern than castration. Use of an additional assay to validate the results may also be required.

It would also be very difficult to draw definitive conclusions about the potential mechanisms by which a screened compound is producing its effect(s) because the presence of an intact HPG axis allows for numerous alternative pathways other direct action on androgen receptors (ARs). The castrated male Hershberger bioassay provides clear, readily interpretable data on (anti-)androgenicity mediated by (ARs). In the weanling version, additional mechanisms, including metabolic inhibition and hypothalamic and/or pituitary regulation of the gonad, and/or indirect effects on the intact HPG axis, are also possible, making the interpretation of the results more complicated. However, when (anti-)androgen action via a non-AR mechanism is suspected, the weanling version of the Hershberger bioassay may be a more appropriate choice.

Another possible confound of the weanling version of the Hershberger bioassay, which was not addressed in the report, is the potential for the administration of testosterone propionate (TP) to advance pubertal onset. Male rats normally undergo pubertal maturation around 45 days of age at which point gonadotropin secretion elevates to adult levels. This is why it is suggested that animals between 21 and 24 days be used in the gonadally intact weanling version of the assay (so that sacrifice will occur a week before pubertal onset). It is possible that the administration of TP (and or the compounds being screened) will "awaken" the HPG axis and accelerate pubertal onset in the test animals, which would then further reduce the sensitivity of the assay. It is difficult to reliably and accurately determine if a male rat is entering puberty. Prepucial separation is one marker that is often used but this measure is not reliable and could be confounded by any of the compounds used in the assay. Timing of pubertal onset is not a concern in the castrate version of the Hershberger bioassay because the animals are already reproductively mature at the time of castration.

1.C. Protocol: In general, the protocol is clearly written and appropriate, however I have a few concerns about some of the details.

1. DIET: The diet should be free of phytoestrogens. As written, there are no restrictions on the diet although each lab is supposed to report which diet was used. The presence of phytoestrogens, even in small quantities, needlessly impairs the sensitivity of the assay and introduces inter-laboratory variability. The Draft OECD Guideline for the Testing of Chemicals (Revised October 20, 2008) suggests that (page 8, line 290), "dietary levels of phytoestrogens should not exceed 350 μg of genistein equivalents/gram of laboratory diet." 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 or at least minimize this potentially problematic source of endocrine disrupting compounds.

2. LITTER EFFECTS: It is not clear from how many litters the group of weaned males will be pulled. The protocol states that up to 6 weanlings can be housed together but it is not clear if all 6 can be pulled from the same litter or should be combined from different litters. Quality of maternal care can impact male sexual development and therefore the potential for a significant litter effect is a concern that should be mitigated by clearly stating that the males must come from multiple (preferably at least 3) litters.

3. TIMING OF ADMINISTRATION: It is not stated at which time of day the compounds are to be administered. Endogenous secretion of androgens has a distinct circadian pattern with levels generally higher in the morning. Therefore the time at which the test compounds are administered should be standardized across labs to ensure that endogenous androgen levels are similar. This is not a major concern in the adult castrate version of the assay because the testes are removed.

1.D. Reproducibility: Inter-laboratory variability and reproducibility is a significant concern. Inter-laboratory variability is unquestionably higher for the weanling version of the Hershberger assay than for the castrated adult version. There was a statistically significant effect of laboratory for nearly every compound and every endpoint tested including the effect of TP. This is particularly worrisome because it is the positive control group. This is likely due to the substantially decreased sensitivity of the assay and the increased technical skill required to dissect and weigh each tissue correctly.

Many of the labs found "marginally insignificant" effects. The Draft OECD Guideline for the Testing of Chemicals (Revised October 20, 2008) states that the study should be repeated when (page 8, line 320), "at least two target tissues were marginally insignificant, i.e. p values between 0.05 and 0.10). Under this requirement a few of the labs (Bayer Crop, Bayer Health, BASF and possibly Korea) would have to repeat the linuron (LIN) test (based on the data presented in Table 6.8 of the Immature Male Rat Hershberger Model Validation Report). Increasing the group size would likely be required to generate enough statistical power to make meaningful conclusions and avoid "marginally insignificant" results. Either way (increasing group size or repeating the assay) would require the sacrifice of more animals than the castrate version, an animal welfare issue that is of greater concern than castration.

The vast majority of compounds screened for endocrine disruption would likely be weak (anti-) androgens (for example, DDE), therefore the reliability and reproducibility of the bioassay for these types of compounds is absolutely critical. The data presented in the report do not sufficiently demonstrate that the weanling version of the Hershberger bioassay can reliably detect the effects of weak (anti-)androgens.

1.E. Bias: A sufficient number of the reference chemicals were tested under code to adequately exclude the risk of bias.

1.F. Performance: The weanling version of the Hershberger bioassay is intended to replace the castrated adult version of the Hershberger bioassay. Chapter 7 of the Immature Male Rat Hershberger Model Validation Report summarized and compared data obtained using each version of the bioassay. Although there are many problems and caveats with this type of comparison (adequately stated in the report) it is an appropriate "first order" method for comparing the efficacy of both assays. CVs for the weanling version were included in the report, and CVs for the castrated adult version were obtained from the Draft OECD Guideline for the Testing of Chemicals (Revised October 20, 2008). Many of the endpoints in the weanling assay had CVs that were higher than the desired maximum. It does not appear that the two versions of the assay were compared in the same laboratory, at the same

time, by the same individuals, which would be the best and most effective way of comparing the performance of each. Therefore it is difficult to determine if/how variables such as strain differences, timing of dose (morning or afternoon), inexperience, diet, or other potential confounds contributed to the outcome differences observed between the two versions.

The sensitivity of the two versions is clearly different, with the castrated version being significantly more sensitive than the weanling version. The weanling version requires a higher dose of TP and the degree to which organ weights change following exposure is clearly smaller. A smaller effect size likely means that larger group sizes are needed to obtain sufficient statistical power. This would mean sacrificing more animals, which defeats the purpose of adopting the weanling version over the castrated version for animal welfare purposes. CVs in the weanling version are generally acceptable but beyond maximum allowable limits in many cases, even within the control group. This likely stems from the technical skill required to properly dissect and weight the weanling organs, which is substantially smaller than the corresponding organs in the adult castrate.

1.G. GLP: It appears that GLP was followed for all studies.

1.H. Peer Review: To my knowledge the peer review is being conducted properly and the protocol for both versions of the Hershberger assay are in the public domain. A few details should be added to the protocol, as described in detail above in 1.C., including the time of day in which the compounds should be administered, the number of litters from which the weanlings should be collected, and the use of a phytoestrogen-free diet.

2. Recommendation: It does not appear that there is sufficient scientific or animal welfare justification for using the weanling version of the Hershberger bioassay in place of the castrated peri-pubertal/adult castrate version. The sensitivity of the weanling assay is demonstrably lower and the degree to which the organ weights change following (anti-)androgen administration is significantly smaller. The magnitude of each effect is also substantially smaller and, by extension, so is the margin for human error when collecting the organs and obtaining the weights. These limitations alone are sufficient justification for rejecting the suggestion that the weanling version is an adequate, appropriate, and effective substitute for the castrate version.

It appears that animal welfare concern over the castration procedure prompted the evaluation of the weanling version as a potential alternative. Unfortunately, the lower sensitivity of the weanling version compounded by the small effect size of each endpoint raises the likelihood that larger group sizes would be needed to overcome inter- and intra-laboratory variability and achieve reliable, accurate and repeatable results. Increasing group sizes would require the sacrifice of more animals, a more substantive animal welfare concern than castration. Pain and discomfort associated with castration is easily managed and routinely done so for millions of laboratory animals, companion animals, and livestock.

The only situation where the weanling version may be advantageous over the castrate version is where an alternative mechanism of action (something other than direct interaction with ARs) is suspected. In this case, using a gonadally intact individual would be of interest because the HPG axis could respond to the exposure. However, the high dose of TP needed for the weanling version would likely suppress HPG function, thereby confounding this potential advantage. So even in this specific situation, although it would be better than the castrate version of the Hershberger assay, it would not likely be sensitive enough to sufficiently identify weakly (anti-)androgenic compounds.

In general, the adult castrate version of the Hershberger bioassay has proven to be a reliable, in vivo method for screening (anti-)androgenic compounds. Although the use of weanlings would eliminate the need for surgical castration, saving time for the technicians and avoiding this animal welfare concern, use of the weanling version introduces other complications and animal welfare concerns. These include increased difficulty in obtaining and weighing the organs correctly (because of their smaller size), reduced overall assay sensitivity (thus increasing the likelihood that more animals will need to be sacrificed), and increased risk of needing to repeat the assay or perform a different assay to confirm the findings. Therefore the weanling version of the Hershberger bioassay is not an adequate substitute for the traditional castrate version.

Dr. Richard M. Sharpe [as-received]

1. Comment on the adequacy of the validation program by stating how well it meets the following validation criteria:

A. The rationale for the test method should be available.
This should include a clear statement of the scientific basis, regulatory purpose and need for the test.

The rationale for the assay, its mechanistic/functional basis and the specific reasons why this additional test is being evaluated are well described, as are specific issues that require consideration in the validation and evaluation exercise.

B. The relationship between the test method's endpoint(s) and the (biological) phenomenon of interest should be described.
This should include a reference to scientific relevance of the effect(s) measured by the test method in terms of their mechanistic (biological) or empirical (correlative) relationship to the specific type of effect/toxicity of interest. Although the relationship may be mechanistic or correlative, test methods with biological relevance to the effect/toxicity being evaluated are preferred.

The reasons for choice of endpoint measurements are well described and sensible and are fully justified, biologically. Performance criteria are established and available and expected working practices and evaluation methods are fully detailed and sound.

C. A detailed protocol for the test method should be available.
The protocol should be sufficiently detailed and should include, e.g., a description of the materials needed, such as specific cell types or construct or animal species that could be used for the test (if applicable), a description of what is measured and how it is measured, a description of how data will be analyzed, decision criteria for evaluation of data and what are the criteria for acceptable test performance.

The test requirements in terms of animals, treatment regimes and routes and standard operating procedures for retrieving endpoint tissues for weighing are well described in the relevant **Draft OECD Guideline for the Testing of Chemicals** document. This is important information for this particular assay as variability in consistency of tissue recovery is of fundamental importance in determining the effective working of the assay; some of the endpoint tissues (SVCG, VP, and to a lesser extent the epididymis) can be fluid-filled with secretions that are androgen-dependent and are thus an integral part of the measured endpoint. Prevention of leakage of these secretions in a consistent manner (by the use of a hemostat) as described in the guidelines is thus vitally important and a key means of reducing measurement errors and variability in weight of the retrieved tissues. It is also important that emphasis is placed on the importance of having the same individual do all of the dissections and weighings whenever possible. The same document tabulates acceptable variation in endpoint measurements (CVs); these are endpoints with naturally high variability and, compounded by variation in dissection/weighing, means that acceptable CVs are set in the 20-40% range for the most useable endpoint tissues. The most useful endpoint (SVCG) based on the evaluation studies is also one of the most variable. Data analysis is described in detail and for participating laboratories will undoubtedly require expert statistical input.

D. The intra- and interlaboratory reproducibility of the test method should be demonstrated.

Data should be available revealing the level of reproducibility and variability within and among laboratories over time. The degree to which biological variability affects the test method reproducibility should be addressed.

The variability of the test is adequately described for participating laboratories and, as would be expected, shows a considerable range with some labs performing better than others in a reasonably consistent way, as is the norm for any assay run in different laboratories. Variability over time was not evaluated at this stage. The use of coded standard samples is an excellent way of objectively evaluating assay performance in a blinded fashion (Tables 6.2, 6.5 and 6.8), although the fact that chosen doses of the test compounds in the earliest comparison were those that induced quite large effects (30mg/kg for flutamide, 100mg/kg for Linuron, 160mg/kg for DDE) perhaps makes this evaluation less of a test of how it would operate in practice with weaker anti-androgens. However, the follow-up blinded assessment using two expected negative compounds (DNP, NP) and a lower dose of flutamide (3mg/kg) gives some measure of reassurance of the utility of the assay as none of the 6 participating laboratories found any significant effect of either of the negative test compounds and all 6 laboratories found significant effects of the flutamide on the weights of androgen-dependent target organs (Table 6.5). A much more convincing evaluation was the 6-laboratory evaluation of coded anti-androgenic chemicals (DDE at 16 and 160mg/kg, Linuron at 10 and 100mg/kg) with results in Table 6.8. This showed that virtually all of the laboratories demonstrated significant effects of the higher doses of both compounds although not convincingly for the lower doses; this evaluation was most robust for SVCG but effects/trends were found for all of the other key endpoint organs (VP, LABC, COW). Biological variability is a big issue and is dealt with in several places in the review, but particularly relevant are the CVs obtained in the intact weanling evaluation studies (Tables in chapters 4 & 5) which show, in general, that the assay works with reasonable performance in this regard, especially when set against the acceptable criteria for each endpoint. This aspect is also covered in response to question 2 below.

E. Demonstration of the test method's performance should be based on the testing of reference chemicals representative of the types of substances for which the test method will be used.

A sufficient number of the reference chemicals should have been tested under code to exclude bias.

The assay review includes assessment in different laboratories (3 or 6, depending on the phase) and testing of a candidate weak androgenic environmental chemical (Trenbolone), plus testing of two candidate weak anti-androgenic environmental chemicals (DDE, Linuron) as well as a positive anti-androgen (flutamide). This is not a huge test of the assay (4-5 anti-androgens would have been better), but as the evaluation has also included a blinded test and included evaluation of negative controls (as mentioned above), and these evaluations worked well, I am convinced that the assay works with reasonable sensitivity and specificity.

F. The performance of the test method should have been evaluated in relation to relevant information from the species of concern, and existing relevant toxicity testing data.

In the case of a substitute test method adequate data should be available to permit a reliable analysis of the performance and comparability of the proposed substitute test method with that of the test it is designed to replace.

The position and relevance of the test in relation to other tests of androgens/anti-androgens in rats is well and accurately described and is embedded in the toxicity literature. The fact that this (modified) assay is based on an assay that has been used by pharmaceutical companies for many years specifically for the evaluation of androgenicity and anti-androgenicity in compounds demonstrates that it has a rock-solid foundation. Its relative utility when compared with existing assays is covered in more detail in the response to Q2 below.

> G. *Ideally, all data supporting the validity of a test method should have been obtained in accordance with the principles of GLP.*
> *Aspects of data collection not performed according to GLP should be clearly identified and their potential impact on the validation status of the test method should be indicated.*

As far as I can tell from the report, all studies were conducted according to GLP.

> H. *All data supporting the assessment of the validity of the test method should be available for expert review.*
> *The detailed test method protocol should be readily available and in the public domain. The data supporting the validity of the test method should be organised and easily accessible to allow for independent review(s), as appropriate. The test method description should be sufficiently detailed to permit an independent laboratory to follow the procedures and generate equivalent data. Benchmarks should be available by which an independent laboratory can itself assess its proper adherence to the protocol.*

The assay review and supporting documents provide sufficient detail for all aspects of the assay, its operation, its methods and analysis to enable its introduction and use by a new laboratory. Benchmark data and statistics are provided that should enable a new laboratory to evaluate its own performance, and I presume that if the assay is approved and adopted then mechanisms will be established to allow for the testing of coded samples or samples for quality control by new laboratories.

2. Given that OECD is considering to apply the same performance criteria for both assays (the same upper CV limits, derived from the castrate version of the assay), is the weanling version of the Hershberger an adequate substitute for the traditional castrate version of the Hershberger assay?

There are three considerations in answering this question: comparative specificity of the two assays, comparative sensitivity of the two assays and comparative variability of endpoints in the two assays (comparison of CVs for the same endpoints). Overall conclusions are then drawn.

Comparative specificity of the intact weanling assay versus the castrate rat assay

The intact weanling assay showed high specificity for androgens, anti-androgens and inactive compounds, although this is based on testing of a very limited number of compounds when compared with the extensive (historical) evaluation that has been undertaken using the castrated rat version of the assay. However, considering the near-identical basis for the two assays, there is no reason to expect that there will be any major difference in specificity of the intact weanling assay in comparison with the castrated rat assay.

Comparative sensitivity of the intact weanling assay versus the castrate rat assay

One of the main reservations expressed about the intact weanling Hershberger screening assay is its lower sensitivity to detect androgenic or anti-androgenic chemicals when compared to the castrated adult or pubertal rat versions of the Hershberger assay (discussed below). This is a pivotal point as the optimal screening assay is one that has the minimum false negative detection rate, and lack of sufficient assay sensitivity would be likely to increase the false negative rate. This raises the issue of what determines assay sensitivity? In the intact weanling assay it will be determined by the combined influences of:

1. The sensitivity of the target (androgen-dependent) organs to androgen action, which in turn will be determined by the number of androgen receptors (AR).
2. The magnitude of response of the target tissues (i.e. the weight range within which the target organ can be changed).
3. For anti-androgens, the level of stimulation by testosterone/other androgens, this representing the summation of exogenously administered testosterone propionate (TP) plus any endogenous androgens.

Item 1 is essentially fixed, although the available evidence suggests that androgen exposure will perhaps upregulate AR expression. Items 2 and 3 are to a large extent inter-related but are discussed separately for clarity.

Item 2 has a wide range for androgenic chemicals but the range for anti-androgenic chemicals is determined by the dose and duration of TP administration. Sufficient TP stimulation needs to have been applied to give a good working range in which anti-androgenic chemicals can effectively antagonize. This was a key goal of the validation/set-up studies in which a TP dose-response curve was undertaken. As with optimizing all assays, a key aim is to ensure that it is ***not*** operated at the upper end (or in excess of the maximally-stimulating dose) of the dose-response curve but rather on the most linear (steepest) part of the curve. In general, experience shows that operating towards the middle of the steepest part of the curve gives the greatest potential for 'unknowns' to push the response up or down steeply and with reasonable sensitivity. This is the situation for an assay that is designed to measure activity in unknowns, whereas in the intact weanling assay for anti-androgens, a change only in one direction (downwards) is being sought for, and therefore the minimum dose (1 mg/kg) of TP that gave close to a maximal stimulation of androgen target organs was used, so as to provide the maximum working range within which test anti-androgenic chemicals could exert their effect.

Though this is reasonable thinking, it does to an extent ignore the issue of assay sensitivity. In essence, the more androgen that a test anti-androgenic chemical is competing with (for binding to a fixed number of AR) at androgen target organs, the smaller an effect it will have and thus the lower the sensitivity of the assay. In practice, in setting up an assay there is inevitably a 'trade off' between assay sensitivity and magnitude of effect because, for example, if the intact weanling assay was operated towards the lower end of the TP dose-response curve then the magnitude of effect that a test anti-androgenic chemical could have would be more limited, which would likely result in a higher CV and greater difficulty in detecting an effect. In many respects, assay sensitivity is usually tailored according to its purpose. For the OECD, two concerns have to be taken into account, namely assay sensitivity and robustness, bearing in mind that the assay needs to be easily useable in many different laboratories, some of which may lack prior experience.

It is clear from Table 4.5 that the 0.8mg TP dose gave nearly the same degree of increase in weight of target organs (and thus the working range for the assay) as did 1mg TP and it was stated towards the end of this section that both doses would be studied further when evaluating the assay using test compounds. In fact only data for 1mg TP is subsequently shown, so I presume that 0.8mg was not tested further. This decision is a little odd as, for the reasons detailed above, use of the 0.8mg TP dose would have been the more logical choice as it could ***only*** make the assay more sensitive compared with 1mg TP without affecting the working range of the assay or increasing CVs (Table 4.5); the increase in sensitivity would probably have been modest but I am surprised at why the 1mg dose was preferred for use.

With only minor exceptions and at every dose of flutamide tested, the castrate rat Hershberger assay showed a notably larger suppressive effect than did the weanling assay, especially for VP and SVCG (Table 7.2). The magnitude of difference was, however, most pronounced for the higher doses of flutamide (3 and 10mg/kg) and was more marginal for the lower doses (0.1 and 0.3 mg/kg). It is these lower doses that are more relevant to the effective screening of weak anti-androgenic environmental compounds. In general, when the sensitivity of the two assays to detect a weak androgenic (Trenbolone) and two weak anti-androgenic (DDE, Linuron) chemicals was compared, the differences found largely reflected expectations based on the flutamide dose-response. Thus, for comparison of the two assay's abilities to detect Linuron (3, 10, 30 or 100mg/kg), there was only minor evidence for higher sensitivity of the castrate versus the intact weanling assay (Table 7.4), and only the top dose of Linuron (100mg/kg) had reasonably consistent significant effects in both assays. A similar comparison for DDE (5, 16, 50, 160mg/kg) provided clearer evidence of higher sensitivity of the castrate assay compared with the weanling assay, as the former detected anti-androgenic effects of 50mg/kg DDE in 1 or more/all of the participating laboratories for all 4 target organs (VP, SVCG, LABC, COWS) (Table 7.5). In contrast, in the weanling assay at this dose, a significant effect was only detected for the most robust endpoint (SVCG), and then only in 2 out of 3 laboratories; no significant effects were detected for the VP, LABC or COWS (Table 7.5). In both assays, significant effects were detected for all 4 organs at the next highest dose (160mg/kg) of DDE (Table 7.5). Despite the sensitivity difference between the two assays, the overall perspective was that either trends or effects of DDE were detected in the intact weanling assay at doses of DDE that caused effects or trends in the intact castrate assay, so the difference in performance between the two assays was a relatively modest one.

One potential reason for differing sensitivity of the two assays is that, in the castrate assay, the anti-androgens are competing against administered doses of 0.2 or 0.4 mg/kg TP whereas in the weanling assay they are competing against an administered dose of 1mg/kg TP (plus any endogenously produced androgens, though this is likely minimal). This is bound to make a difference, though it is obviously not the full explanation.

Comparative variability of the intact weanling assay versus the castrate rat assay

For this comparison, I used data for the castrate Hershberger assay published by Owens et al (2007) in Environmental Health Perspectives (115: 671-678). Overall, there are no major differences in CVs for the various endpoints in the two assays. Therefore, at least for this rather limited comparison, I do not see endpoint variation in the intact weanling assay as being a limiting factor in its operation and utility. Arguably the most important evaluation of the intact weanling assay performance was in the 6-laboratory evaluation of coded anti-androgens (DDE and LIN at two doses each). This showed that measurement variation outside of the pre-set maximum allowable CV occurred sporadically rather than consistently (Table 6.10). Thus allowable CVs were exceeded once for VP (DDE 16mg/kg), once for SVCG (DDE 160mg/kg), three times for LABC (one at DDE 16mg/kg, two at DDE 160mg/kg) and 4 times for COWS (one at DDE 16mg/kg, two at DDE 160mg/kg, one at LIN 10mg/kg). The important messages from this evaluation are (1) the sporadic nature of these excesses means that it is not directly assay-related (i.e. there is not something intrinsically wrong with how the assay works); (2) the excesses occurred as frequently with higher doses of anti-androgens, which caused clear anti-androgenic effects, as they did at lower doses where effects were marginal; (3) the target organ identified from the evaluation studies as providing the most sensitive and robust response to both androgens and anti-androgens in the intact weanling assay, namely the SVCG, was the best performing endpoint (along with VP) in terms of meeting acceptable CVs (Table 6.10). Based on comparison of CVs, there is no reason to reckon that the weanling assay will perform any differently than the castrated rat assay.

Overall conclusions

The intact weanling assay is fit for purpose and can provide an adequate substitute for the castrated rat assay. The latter is slightly more sensitive and provides a wider working range of target organ weights, but its overall operation is not markedly different from that of the intact weanling assay. Arguably, when large numbers of compounds are screened, it is likely that the intact weanling assay may not detect some weak anti-androgens that the castrate adult assay does detect, but this is likely to be infrequent. This possibility is minimized by the use of multiple evaluation endpoints, and this is an inherent strength of the intact weanling assay, just as it is for the castrated rat assay.

It is stated in the evaluation report that one advantage of the intact weanling assay is that having an intact hypothalamic-pituitary-testicular (HPT) axis, means that it may detect a wider range of anti-androgenic compounds than the castrated rat assay. This remains only a theoretical possibility as there are no demonstrations that this is indeed the case – in my opinion, the HPT axis will be essentially shut down due to the TP treatment. One present drawback with the intact weanling assay is that there is very limited history in its use, whereas there is a huge depth of experience with the castrated rat assay.

Dr. Chris E. Talsness [as-received]

INDEPENDENT PEER REVIEW OF THE WEANLING VERSION OF THE HERSHBERGER ASSAY

Chris E. Talsness

1. Comment on the adequacy of the validation program by stating how well it meets the following validation criteria:

A. The rationale for the test method should be available.
This should include a clear statement of the scientific basis, regulatory purpose and need for the test.

The **scientific basis** for the test is founded and has been well documented and employed in assays since the 1930's. In addition, extensive validation work has successively been performed on the castrated version of the assay. The test evaluates the ability of a substance to increase or decrease the weight of androgen-dependent tissues of the pre-pubertal male. At the time of the test, starting between PND 21-24 and ending between PND 30-33, androgen receptors are present on the tissues evaluated and the serum concentration of testosterone is relatively low and begins to slowly increase over time with a peak occurring between 50 and 60 days of age. Due to the relatively low testosterone concentrations, the androgen dependent tissues can respond to exogenous androgens and anti-androgen effects can be demonstrated with co-administration of testosterone propionate in comparison with testosterone propionate alone. 5α reductase inhibitors can be theoretically detected based on a differentiated response among the tissues due to their differing relative dependency on 5 α reductase conversion, although validation of this was not performed, *e.g.*, finasteride. The hypothalamic-pituitary-gonadal axis is functional in this model and indirect effects on this axis can theoretically be detected although validation for this was not performed*, e.g.*, GnRH agonists or antagonists.

The **regulatory purpose** of this assay is to serve as a Level 3 test which includes *"in vivo* assays providing data about single endocrine mechanisms and effects". Although the weanling assay theoretically covers a larger scope of investigation than the castrated version, only information regarding (anti) androgenicity in the broad sense is obtained. Screens should be sensitive methods with the emphasis placed on reducing the number of false negatives while possibly increasing the number of false positives. The weanling version is clearly less sensitive than the castrated version and appears to have similar specificity as the castrated version. The doses required for identification of substances were higher in the weanling version than the castrated one, however, each chemical was eventually identified at the highest dose as in the castrated version*, i.e.,* the resulting information obtained from the two tests was the same. This test guideline, therefore, could be used for hazard characterization. Calculation of reference doses should be performed with data derived from other *in vivo* tests investigating other types of endpoints.

The **need** for the test is to reduce the number of chemicals identified in level 2 testing which will require further investigation using more elaborate protocols to evaluate possible endocrine-related effects and to address welfare concerns questioning the need to use a castrated model to achieve this aim.
In vitro tests do not account for the influences of absorption, distribution, metabolization or elimination (ADME) and it is theoretically possible that "positive compounds" from *in vitro* screens will be rendered "inactive" in the *in vivo* model, thereby lowering the number

of chemicals indicated for further evaluation of endocrine effects (level 4). It is just as theoretically possible that negative compounds in the *in vitro* testing phase could be positive *in vivo* due to the influences of ADME which leads to the question whether the Hershberger Bioassay should be reserved instead for chemicals which are negative in level 2.

Evaluation of the test as an alternative protocol to the castrated model version is justified.

B. The relationship between the test method's endpoint(s) and the (biological) phenomenon of interest should be described.
This should include a reference to scientific relevance of the effect(s) measured by the test method in terms of their mechanistic (biological) or empirical (correlative) relationship to the specific type of effect/toxicity of interest. Although the relationship may be mechanistic or correlative, test methods with biological relevance to the effect/toxicity being evaluated are preferred.

The required endpoints include: daily body weight and clinical observations and weights of Cowper's glands, levator ani-bulbocavernosus muscle complex, seminal vesicles with coagulating glands and their fluids and ventral prostate. Optional assessments include daily food consumption, weights of paired testes, paired epididymides, liver, paired kidneys and paired adrenal glands and serum LH, testosterone, T4 and T3.

Evaluation of clinical appearance and daily body weight are reasonable endpoints to provide relevant information to assess possible overt toxicity and allow for adjustment of dose during a period of rapid growth. In addition, body weight allows analyses to be performed regarding possible influences of body weight on weights of accessory sex tissues, the main endpoints of the assay. Change in the weights of the accessory sex tissues is a biologically sound expectation to exposure to exogenous androgens or co-exposure to androgen/anti-androgen because these tissues are dependent on testosterone and 5 α dihydrotestosterone throughout puberty and adulthood for maturation and function. The weanling model is characterized by 1- low baseline serum testosterone concentrations allowing the detection of a response to androgen agonists as the tissues are not in a state of maximal stimulation, 2- the presence of androgen receptors on the accessory sex tissues allowing a response and 3- steady physiological growth of these tissues with any dramatic changes typically occurring after the timeframe of the study period.

Optional Endpoints: Paired weights of the testes and epididymides

- Androgen agonists
Detection of statistically significant changes in the weights of the epididymides and the testes required a higher dose of TP compared to the mandatory accessory sex tissues. In addition, the relative response of the epididymides is also lower perhaps making detection with this organ more difficult.

Only one laboratory detected changes in the epididymides in response to TREN while the testes data (dose and response) supported the effects observed in the mandatory accessory sex tissues.

- Androgen antagonists
A higher dose of FLU was required to detect a statistically significant response in paired testes weight and statistically significant changes were not observed in response to LIN or DDE
Epididymides responded at to FLU at the same dose as the mandatory accessory sex tissues and did not respond to LIN or DDE at all doses tested.

In general, the opportunity to evaluate these two organs did not increase the sensitivity of the assay and they proved to be "weak detectors" of androgen antagonists, but changes in weight (although not statistically significant) may be useful as further evidence when equivocal results are obtained with the mandatory accessory sex tissues.

C. *A detailed protocol for the test method should be available.*
The protocol should be sufficiently detailed and should include, *e.g.,* a description of the materials needed, such as specific cell types or construct or animal species that could be used for the test (if applicable), a description of what is measured and how it is measured, a description of how data will be analyzed, decision criteria for evaluation of data and what are the criteria for acceptable test performance.

The protocol is fairly detailed and provides significant information to perform the assay successfully. Information regarding excision of the sex accessory tissues is provided in the guidelines for the castrated version and a description for the testes and epididymides should be included here. Guidance should be included in terms of the number of doses required for test substances and the information that greater doses most likely have to be employed in the weanling mode than in the castrated version. Criteria for acceptable test performance need to be determined and/or included in the document (comment based on report identified above). Additional discussion regarding acceptable standard curves (weanling version has a flatter dose response curve for TP) with reference substances and that this should be repeated at appropriate intervals to ensure proper test performance. Finally, more information regarding the interpretation of data is required, *e.g.,* what constitutes a positive response when only some of the tissues exhibit changes that are statistically significant.

D. *The intra- and interlaboratory reproducibility of the test method should be demonstrated.*
Data should be available revealing the level of reproducibility and variability within and among laboratories over time. The degree to which biological variability affects the test method reproducibility should be addressed.

Intra-laboratory reproducibility

Three laboratories (Syngenta, Korea, Canada) repeated the following combinations: Vehicle *vs* 1 mg/kg TP, 1mg/kg TP *vs*. 1mg/kg TP + 3 mg/kg FLU, 1mg/kg TP *vs*. 1mg/kg TP + 100 mg/kg LIN, and 1mg/kg TP *vs* 1mg/lg TP+160 mg/kg DDE. The changes in weights of the mandatory accessory sex tissues obtained in the two experiments were roughly compared for each laboratory for each of these combinations. The changes in tissue weights achieved by each laboratory for the individual tissues were roughly similar over time.

Exceptions were:

Vehicle *vs* 1 mg/kg TP
VP for Syngenta

1mg/kg TP *vs*. 1mg/kg TP + 3 mg/kg FLU
VP and COW for Syngenta
VP and SVCG for Korea

1mg/kg TP *vs*. 1mg/kg TP + 100 mg/kg LIN
SVCG for Korea and Syngenta

1mg/kg TP *vs* 1mg/lg TP+160 mg/kg DDE
VP for Syngenta and Canada
SVCG for Syngenta, Korea and Canada

In general, the laboratories were able to reproduce their results between the two experiments.

Inter-laboratory reproducibility
There is significant influence of the laboratory on the all of the organ weights regardless of the compound tested. However, all laboratories were able to achieve the same results at the highest doses tested. Where standard curves are available, not all laboratories were able to detect statistically significant changes in organ weights at the same doses (lower) tested.

It is expected that the biological variability in the weanling model would be higher than in the castrated model due to the low level of physiological testosterone and presence of an intact hypothalamic-pituitary-gonadal axis.

E. *Demonstration of the test method's performance should be based on the testing of reference chemicals representative of the types of substances for which the test method will be used.* A sufficient number of the reference chemicals should have been tested under code to exclude bias.

The following chemicals were coded and evaluated in Phase 3: TP (agonist), FLU, DDE, LIN (antagonists), and DNP and NP (negative chemicals).
A coded agonist was not tested *per se* as some laboratories did not perform a vehicle and statistical analyses between TP and vehicle was not performed. It is expected that more chemicals exhibit anti-androgenic effects than androgenic and priority was placed on the correct reference chemicals.

<u>DNP and NP</u>
None of the laboratories detected statistically significant changes in organ weights following exposure to the coded negative chemicals and all demonstrated significant changes with the positive control.

<u>DDE</u>
Androgen antagonism was not detected by any laboratory for the low dose of coded DDE.
Androgen antagonism was detected for all mandatory tissues by all laboratories for only the high dose of coded DDE.

<u>Linuron</u>
No laboratory detected androgen antagonism with the low dose of coded linuron.
Only one laboratory detected androgen antagonism with the high dose of coded linuron in the VP.
5/6 detected androgen antagonism with the high dose of coded linuron in the SVCG.
3/6 detected androgen antagonism with the high dose of coded linuron in the LABC.
1/6 detected androgen antagonism with the high dose of coded linuron in the COW.
On a per laboratory basis,
1 laboratory detected 4 statistically significant organ weight changes for the mandatory tissues indicative of antagonism

2 laboratories detected 2 statistically significant organ weight changes for the mandatory tissues indicative of antagonism

1 laboratories detected 1 statistically significant organ weight change for the mandatory tissues indicative of antagonism

2 laboratories did not detect any statistically significant organ weight changes for the mandatory tissues indicative of antagonism

Depending on the criteria employed (*e.g*., detection in at least 2 tissues), 50% of the laboratories did not identify linuron correctly.

The false positive rate of the two chemicals tested was 0 and the false negative rate for detection of weak anti-androgens may need improvement particularly as this test is to function as a screen where sensitivity is a priority over specificity.

In the phase 3 testing for the castrated model (http://www.oecd.org/dataoecd/49/34/37479136.pdf),
6/10 laboratories detected 5 statistically significant organ weight changes for the mandatory tissues indicative of antagonism
3/10 laboratories detected 4 statistically significant organ weight changes for the mandatory tissues indicative of antagonism
1/10 laboratories detected 3 statistically significant organ weight changes for the mandatory tissues indicative of antagonism

Depending on the criteria employed (*e.g*., detection in at least 2 tissues), all laboratories were able to correctly identify linuron at the 100mg/kg/d dose indicating better sensitivity for the castrated version.

F. The performance of the test method should have been evaluated in relation to relevant information from the species of concern, and existing relevant toxicity testing data.

In the case of a substitute test method adequate data should be available to permit a reliable analysis of the performance and comparability of the proposed substitute test method with that of the test it is designed to replace.

Evaluation of the weanling version of the Hershberger Bioassy was conducted in a similar fashion (same doses and methodical procedure) to the castrated version allowing for comparison of the two methods. In general, sufficient data is available to allow comparison of the two methods. Analysis of a 5α-reductase inhibitor to assess detection of compounds with this mode of action would have allowed direct comparison of the three types of substances evaluated in the castrated version. Finally, presentation of GP data from the castrated version would have been helpful of the reviewer to assess whether pertinent information was lost with loss of this endpoint.

G. Ideally, all data supporting the validity of a test method should have been obtained in accordance with the principles of GLP.

Aspects of data collection not performed according to GLP should be clearly identified and their potential impact on the validation status of the test method should be indicated.

No comment.

H. All data supporting the assessment of the validity of the test method should be available for expert review.

The detailed test method protocol should be readily available and in the public domain. The data supporting the validity of the test method should be organised and easily accessible to allow for independent review(s), as appropriate. The test method description should be sufficiently detailed to

permit an independent laboratory to follow the procedures and generate equivalent data. Benchmarks should be available by which an independent laboratory can itself assess its proper adherence to the protocol.

The data for this review was presented in a logical and organized fashion. The review could be conducted efficiently and information was easy to find. Suggestions for additional information for the test method have been described above. .

2. Given that OECD is considering to apply the same performance criteria for both assays (the same upper CV limits, derived from the castrate version of the assay), is the weanling version of the Hershberger an adequate substitute for the traditional castrate version of the Hershberger assay?

CVs for the data presented for the **weanling** model
-TP: standard curve (Table 4.4)
-Tren (Table 5.4)
The coefficients of variation of the data from all the laboratories used to evaluate Tren and to generate the standard curve with TP were below the maximum allowable CV's indicating adequate performance off the assay. According to this standard, the three laboratories were able to successfully complete this portion of the experiment.

The following indicates where the maximum CV values were surpassed.
-TP+FLU (Table 4.9)
The CVs were exceeded by one laboratory for the COW

-LIN (Table 5.6)
The CVs were exceeded by one laboratory for the VP and the LABC and by one laboratory for the COW.

-TP + DDE (Table 5.10)
The CVs were exceeded by 2/3 laboratories for the LABC.

-TP + FLU (Table 6.7)
The CVs were exceeded by 3/6 laboratories for the COW.

Regardless of substance tested, the CVs for the SVCG were never over the maximum allowed.

CVs for the data presented for the **castrate** model

Table 11 in the phase 3 report for the castrate model indicates that for evaluation of agonists, the CV was surpassed for
-4/10 laboratories for the VP
-4/10 laboratories for the SVCG
-2/10 laboratories for the COW

Table 15 in the phase 3 report for the castrate model indicated that none of the 10 laboratories exceeded the maximum CV for the antagonist data for any of the mandatory accessory sex tissues.

The CVs for the GP and LABC were never over the maximum allowed.

Even though tissue dissection may be more challenging in the weanling model, it appears to be technically feasible as in most cases the laboratories were able to achieve CVs under the

maximum limit. A hundred percent success rate was also not achieved in the castrated version although the animals are bigger and the tissues have been theoretically been primed to react in a more pronounced fashion. In this respect, the castrated version does not provide an advantage over the weanling model.

Sensitivity

Comparison of the weanling and castrated versions of the assay in terms of changes in tissue weight in response to exposure to testosterone propionate (Table 7.1) indicates that the castrated version of the Hershberger bioassay is more sensitive. Statistically significant changes in tissue weights were observed by all laboratories in all tissues starting at 0.2 mg TP /kg/d for the castrated version and at 0.8 mg TP/kg/d for the weanling version. It is also apparent that the dose response curve for the weanling version is much flatter. The relative changes in tissue weights are much smaller and probably accounts for the increased difficulty in detecting statistically significant changes in weight.

Comparison of the weanling and castrated versions of the assay in terms of changes in tissue weight in response to exposure to testosterone propionate and flutamide (Table 7.2) indicates that the castrated version of the Hershberger bioassay is more sensitive. Statistically significant changes in tissue weights were observed by all laboratories in all tissues starting at 1 mg FLU /kg/d for the castrated version and at 10 mg FLU/kg/d for the weanling version.

Comparison of the weanling and castrated versions of the assay in terms of changes in tissue weight in response to exposure to trenbolone (Table 7.3) indicates that the castrated version of the Hershberger bioassay is more sensitive. Statistically significant changes in tissue weights were observed by all laboratories in all tissues at the highest dose (40 mg TREN /kg/d for the castrated version and only 2/4 tissues were determined to be statistically significant by all laboratories at the highest dose of 40 mg TREN /kg/d for the weanling version.

Comparison of the weanling and castrated versions of the assay in terms of changes in tissue weight in response to exposure to testosterone propionate and linuron (Table 7.4) indicates that neither version of the Hershberger bioassay performed particularly well in detecting antagonistic action with this substance. At the highest dose tested, all of the laboratories were able to detect a statistically significant change in tissue weight for only one tissue in both versions of the bioassay. However, the comparison in Table 7.4 is somewhat misleading as it is more important to know how a lab would identify a substance as to how many labs observed a change in a specific tissue weight. Looking at the original data on a laboratory basis reveals that ¾ labs would correctly identify linuron and one laboratory had an indication in one organ using the castrate model as

1 lab detected changes in all five tissues
2 labs detected changes in four tissues
1 lab detected changes in one tissue

In the weanling model:

2 labs detected changes in three out of four tissues
1 lab detected changes in two out of four tissues

Looking at the data on a per laboratory basis indicates that both versions were able to identify linuron.

Comparison of the weanling and castrated versions of the assay in terms of changes in tissue weight in response to exposure to testosterone propionate and DDE (Table 7.5) indicates that all laboratories detected a statistically significant change in tissue weight for two of the tissues at the dose of 50 mg DDE/kg/d using the castrated version of the Hershberger bioassay while a statistically significant change was not found by all of the laboratories for any of the tissue weights at this dose. All laboratories observed statistically significant changes for all tissues at the highest dose regardless of method. Again, the castrated version appears to be more sensitive.

In summary, there are examples in both bioassays where the maximum coefficients of variation were exceeded. In this respect, the weanling version is an adequate substitute for the castrated version. Interestingly, CVs were exceeded with testing of antagonists, but not agonists in the weanling version and the opposite is true for the castrated version.
Since it is expected that there are more environmental antagonists than agonists, this may represent a possible advantage for using the castrated version of the bioassay. Based on the limited data, a preliminary deliberation is that it may be prudent to using the castrated version of the bioassay for the cases where the *in vitro* data indicate antagonist androgen action and reserve the weanling when agonist activity is expected.

Another advantage of the castrated version of the bioassay is that significant changes in tissue weights were detected at lower doses of the test compounds than in the weanling version. The increased sensitivity is probably due to the greater relative changes in organ weight. Higher doses would have to be employed in the weanling version of the bioassay which may increase the number of false positives due to non specific action.

The advantages of the weanling version include an intact biological system and avoidance of the surgical castration procedure.

Although the castrated version is more sensitive, there is no example in the data provided where the compound was not eventually identified in the weanling version as in the castrated version and, therefore, the weanling version could be used as a substitute for the castrated version with the caveat that higher doses may have to be employed. The choice of which version to employ could be based on available *in vitro* data.