Report of the Peer Review Panel for the Uterotrophic Bioassay

Summary

1. The peer review panel (PRP) was constituted in September 2003, to provide a review of the validation process for the Uterotrophic Bioassay, to evaluate the data collected, and to answer specific questions posed to the Panel in the charge provided by the sponsoring organization, the Organization for Economic Cooperation and Development (the OECD). The Panel held several teleconferences, and each Panel member submitted written answers to the charge questions to the Secretariat prior to each teleconference. This report presents the combined PRP responses to each of these charge questions.

2. The Peer Review Panel was requested to report their views on the validation process for the Uterotrophic Assay to the Validation Management Group-mammalian (VMG-mammalian) responsible for overseeing the validation process, and then to the Endocrine Disruptor Testing and Assessment Taskforce (EDTA), and the Working Group of the National Coordinators of the Test Guidelines Programme (WNT). Taking this peer review report into consideration, the EDTA and WNT will recommend any further activities on this OECD project, including the process for the development of a Test Guideline.

3. Regarding the overall validation exercise, the final conclusions and the views of the PRP are divided into broad groups and there were considerable differences expressed regarding the various components of the project. The PRP was unable to reach consensus on the issue of the validation status of the uterotrophic assay, and the differences in opinion between PRP members were significant. Some members considered the Uterotrophic Bioassay to be validated for the intended purpose of the assay, other members considered that further data, including on negative substances, was necessary to reach a decision on the validation status of the assay, whilst other members considered the efforts to date were not sufficient to validate the test method but to only be sufficient as a pre-validation study. The difficulty in reconciling such differences should be considered when reading the individual responses to the charge questions that follow.

Background

4. The National Coordinators (WNT) agreed to establish a special activity to address the issue of endocrine disruption and develop new Test Guidelines as appropriate. The responsible body was the Task Force for Endocrine Disruptor Testing and Assessment (EDTA). The EDTA agreed to initially pursue efforts to develop and validate Test Guidelines for the uterotrophic assay and the Hershberger assay, and to evaluate enhancements to the current Test Guideline 407. A single Validation Management Group (VMG) was established to manage these projects. Subsequently, the EDTA has begun activities concerning ecotoxicity testing and *in vitro* or non-animal testing also related to the endocrine disruption issue. As a result, the original VMG is now divided into the VMG for mammalian effects assessment (VMG-mammalian), the VMG for ecotoxicity testing (VMG-eco) and the VMG for non-animal testing (VMG-NA) to manage the diverse activities and work loads.

5. The VMG-mammalian managed the uterotrophic activity, the first step of which was to evaluate and propose the standardization of protocols to address the use of both the immature female rat and ovariectomized female rat, as well as different routes of administration. The VMG-mammalian designed and conducted an initial phase (Phase-1) with the potent oestrogen compound, ethinyl oestradiol, to demonstrate the transferability and reproducibility of the protocols with studies conducted in some 20 laboratories. The VMG-mammalian then initiated Phase-2 with weak oestrogen agonists both in dose response studies and in studies where these same chemicals were coded. These extensive

data sets were analyzed by independent statisticians and provided to the PRP, and many of the data have been published in the scientific literature.

6. The charge of the PRP was to report on the biological and toxicological relevance of the assay, the adequacy of the protocol, the extensive set of data generated in Phases-1 and -2, and the analyses of that data. The PRP was also asked to determine whether the assay was 'validated' to meet the requirements for development of a test guideline based on this test method.

7. To facilitate this report, a series of more than 30 questions were posed to the PRP and the responses to the questions (attached) reflect the individual views of the panel. The more detailed discussions on these issues are provided in the individual reports of the teleconferences. A summary of the views of the Panel members is provided below. Where agreement between the Panel members was reached, this has been indicated. Where agreement was not reached, the text below provides a short summary of the differing views of the panel members on specific issues.

Summary of Panel Responses to Issues

8. The Peer Review Panel agreed that the biology of the test system was appropriate, as the rat uterus is a biologically relevant test system for detecting oestrogen-like biological effects *in vivo*. A majority of the Panel members agreed that the assay was adequate to test effects of oestrogen agonists for the purposes of the assay (which is designed to be used as an *in vivo* screening assay), while recognizing that there are a number of details that need to be clarified in a finalised test method. For example, instructions are necessary for the specifics on dose-setting (including what constitutes a maximum tolerated dose to avoid animal pain and suffering, and if there is to be a limit dose as in most test guidelines); the appropriate statistics for the test method; allowable limits for phytoestrogens in feed; criteria for accepting data as high quality; and, importantly, criteria for determining if a compound is positive or negative. Notwithstanding these issues, this subgroup of PRP members supported the usefulness of the test and agreed that validation is achievable for the purpose of the test, possibly with the addition of some further information on the performance of the assay with other substances.

9. Some members of the PRP expressed the need for additional information on negative compounds, due to the need for negative compounds, in addition to positives, to fully assess the assay in terms of specificity and sensitivity. Some of the PRP members indicated that such information could be obtained in a retrospective manner, from existing studies, and this information would allow further decisions to be made regarding the usefulness of the assay to screen for oestrogen antagonists.

10. Another group of the PRP members believed that the shortcomings in the current validation effort are significant and suggested that the current test should be considered to be in a 'pre-validation stage' only. That is, in the opinion of these PRP members, a validation process is yet to properly begin. This group emphasized the need to include, and to test the response of the Uterotrophic Bioassay with, additional negative substances, and recommended the need for additional metabolically-enhanced positive substances be included where possible. One view of some PRP members and observers of this group is that the testing of negatives was insufficient and would prevent the validation effort from moving from this 'pre-validation' phase.

11. Some PRP members indicated that the assay should meet the validation requirements of regional or national bodies such as ECVAM and ICCVAM, which specialise in validation of *in vitro* and other alternative methods. Some members also stated that the assay, while detecting an increase in uterine weight, did not necessarily confirm that such an effect was attributable to oestrogen agonists, and that other tests (such as receptor binding) should be utilised in conjunction with the uterotrophic assay. This would allow the assay to be used as a screen for detection of oestrogenic effects. On this issue, some

panel members stressed that the validity of any bioassay must be assessed within the context of its purpose and the conditions of its intended use. Thus, if the purpose of the uterotrophic assay is to flag a chemical for further evaluation, and will not be used to label the chemical as having a certain type of activity or mechanism of action, the inability of the assay to distinguish between substances that increase uterine weight via oestrogenic agonism versus other, as yet undefined, pathways is not a basis for rejecting the validated status of the assay.

12. Several PRP members stated that the validation of *in vitro* assays should be a priority for the OECD, as these alternatives promise lower costs and would reduce the number of animals used in the Uterotrophic Bioassay. Other PRP members identified the regulatory needs of OECD member countries for this test method as a rationale for completing the validation of the assay prior to other validation activities being undertaken.

Recommendations

13. The Peer Review Panel agrees that this report provides a summary of their views on the status of the validation of the uterotrophic assay, as detailed in the responses to the questions posed to the Panel members in the form, and based on the information on the validation exercise provided to the Peer Review Panel.

14. The report of the Peer Review Panel, along with the information developed on the validation of the uterotrophic bioassay, should form the basis for decisions on whether the validation exercise meets the OECD principles for validation for development of this test method into an OECD Test Guideline. In this consideration, the OECD should note the various views of members of the PRP. The PRP recommends that the OECD consider the PRP report, along with the validation information, to decide on additional work needed to finalize the validation exercise for the purposes of developing an OECD Test Guideline.

Summary of PRP Responses to Individual Charge Questions

1) Is the choice of the rodent uterus, specifically in this case the rat uterus, biologically relevant for the detection of oestrogen agonists and antagonists in vivo?

15. The Panel found that the rodent (specifically the rat) uterus is biologically relevant for the detection of oestrogen agonists and antagonists, as the uterus has a clear normal response to estradiol (cell growth + fluid inbibition + other trophic responses = weight gain). The Panel finds that the Uterotrophic Bioassay alone is not sufficient to determine and conclude that a compound is oestrogenic. Aromatizable androgens will clearly induce uterine weight increases in the immature animal, and any steroid due to its structural similarity should eventually bind the estrogen receptor as these receptors are evolutionarily related. There are questions about the assay's sensitivity, but the biological relevance of the assay for oestrogenic chemicals was agreed.

2) Is the choice of the rodent uterus, specifically in this case the rat uterus, mechanistically adequate and sensitive for the detection of oestrogen agonists and antagonists in vivo?

16. The Panel found that there is "adequate" mechanistic information for nuclear oestrogen receptor involvement in the trophic response of the uterus. The specific mechanism of some parts of this response (water imbibition) are less well-known than, say, the cell proliferation component of this trophism.

17. The question of sensitivity is more complicated. The test can identify weak oestrogen agonists although the inter-laboratory variability may be an issue. For example, the lowest significantly effective dose for Bisphenol A varied between 10 mg/kg/d and 600 mg/kg/d (Table 9b, Part 2, pg 29) for Protocol B in the immature version, whilst in Protocol C with the OVX animals all labs achieved significance at 100 mg/kg/d. Thus, while the test was capable of identifying an increase in uterine weight in multiple laboratories, there were differences noted for the lowest effective dose for some test protocols. To date it is not clear whether the lowest effective dose will become an important output from this test, or merely the fact that a weight increase of significance was registered at one of the tested doses. Some Panel members expressed the view that if the uterotrophic assay is not intended to be used to define a no-observable adverse effect level (NOAEL), concerns about inter-laboratory variability regarding selection of the lowest effective dose of a substance are not an issue in the context of safety assessment and regulatory use. The issue of sensitivity of rat uterus vs. human uterus or other human oestrogenic responses is clearly important, but was outside the purview of this test method development.

3) Is the choice of the rodent uterus, specifically in this case the rat uterus, toxicologically an appropriate choice for the detection of oestrogen agonists and antagonists *in vivo*?

18. The Panel agreed that the rat is, toxicologically, an appropriate choice for the in vivo component of a set of assays that will identify oestrogen agonists and antagonists. The rat was felt to be more useful than the mouse due to its reduced variability across strains, and the fact that there is a huge historical database stemming from the use of the rat in reproductive assays. Any other species is larger and more expensive, and further animal welfare considerations come into play.

19. The fact that androgens also cause a uterotrophic response compels the use of some additional means of determining whether the response is oestrogenic or androgenic.

20. Additionally, the data generated by the VMG so far for oestrogen antagonists are sufficient to show proof of concept, which is quite important, but they are insufficient to determine how often the assay would correctly detect antagonists since just one single antiestrogen was included in this validation exercise. The Panel identified this specificity issue as one of the concerns that should be addressed in the future work on this assay. One observer felt that this issue of specificity was not addressed sufficiently for any conclusion to be reached; the rest of the Panel agreed that future work could address this issue.

4) Is the choice of the rodent uterus, specifically in this case the rat uterus, consistent with the use of animals to obtain *in vivo* hazard information for human hazard assessment?

21. The Panel agreed that the use of animals is consistent with toxicological principles and the currently-accepted state-of-the-art for conducting toxicological studies. Animals have important homeostatic functions not recapitulated in cell culture, and the use of animals to model the response of humans and wild animal populations is appropriate and necessary. The Panel found that positive results from a uterotrophic assay, alone, is insufficient to characterize the hazard of a compound as "oestrogenic", and that regulatory action should not be based on data from the Uterotrophic Bioassay alone. Uterotrophic Bioassay data should be considered as only one part of a tiered testing strategy, which leads to increasingly comprehensive (and expensive) tests which will define the scope of a chemical's actions. In particular, only from such a comprehensive data set can a risk assessment be conducted. A number of PRP members felt that priority should be given to the development of in vitro alternative tests for this endpoint rather than the in vivo assay.

5) Are the test method and protocols described in sufficient detail in the Submission Package, including the purpose of the test, endpoints, protocol parameters, and acceptable variations among the protocols?

22. The test method as described in its most advanced form (Phase 2, Annex 2) describes what is to be measured, the age of the animals, that records should be kept of possible confounding parameters, how to excise the uterus, and the ways in which the weights are to be collected. Panel members suggested a small number of items they considered important to record, including specification of a vehicle such as tocopherol-stripped corn oil and an acceptable weight range for the animals.

23. However, the Phase-2 Annex 2 text is effectively a model protocol for a specific study meant to evaluate only the inter-laboratory variation using prescribed doses ("All test substances will be tested at the doses specified"). There is room for some further refinement for addressing particular substances or conditions (e.g., the specific method to be used for solubilization; how to choose the doses to use; the specific methods for handling high- vs low-phytoestrogen-content diets; statistics). So some members of the Panel considered this to be the method for a specific study, and not a final propocol per se.

24. To make this test method suitable for development into a final Test Guideline, a number of these issues should be clarified:
    a) The statistical tests to be used in the final version of the protocol are not specified ("The OECD VMG will determine the statistical procedures to be used in the evaluation of data taking into account expert statistical advice."), and this critical component needs to be detailed before the protocol can be considered final.
    b) Setting doses is a significant source of variability in testing, and guidance should be provided in the test guideline to enable laboratories to select appropriate doses for their specific tests.
    c) There is no specification of when a test will be considered positive. It is not specified whether a positive result be considered an absolute increase over controls (say, any weight increase more than 20 mg), a percentage increase over controls, or any statistical increase vs. concurrent controls.

d) The protocols for the validation program specified the routes of administration, and the final test guideline must give guidance for selection of the appropriate route of administration. The appropriateness of one route will vary depending on the use of the chemical and how humans are exposed.

6) Are the protocols used to generate the supporting submission data complete and adequate in detail for a laboratory to conduct the study, including a description of the material and equipment needed to conduct the test?

25. The Panel agreed that the study protocol that was sent out to the dose-response labs was thorough with regard to the equipment and material needed to conduct the test and similar information will be needed for the Test Guideline. There was an appreciation that the performance of this test will require personnel trained in animal in vivo and necropsy procedures, for example, the crucial process of weighing the uterus is highly dependent on the isolation procedure. This important variable may not be effectively controllable unless there are qualification procedures for laboratories, but this is a factor for consideration in the conduct of all *in vivo* animal tests.

7) Are the protocols used to generate the supporting submission data complete and adequate in detail for a laboratory to conduct the study, including a description of what is measured and how the data are used to identify positive and negative results?

26. The Panel agreed that there is a clear description of what is to be measured, and how this is to be done. The study-specific protocols that were evaluated described the animal conduct of an Uterotrophic Bioassay, but there were some views that the protocol left unaddressed how the data were to be handled or interpreted, and that this will be critical as the protocol is finalized. The Panel noted that there is no guidance given on how to evaluate the results, and when to call a compound a positive, and whether compounds should be described as strongly or weakly positive, and where that division is between these descriptions. Some members expressed the need for a prediction model similar to that used by ECVAM for in vitro studies. Some Panel members stated that the Test Guideline should specify the process for interpretation of results, such as by statistical analysis or by providing a clear description of what constitutes a positive result. Conversely, other views included that the detailed description of the statistical method in the reports is sufficient for the interpretation of the results, and that the use of prediction models is for validation studies, not protocols.

8) Are the protocols used to generate the supporting submission data complete and adequate in detail for a laboratory to conduct the study including the appropriate provisions for the use of reference control chemicals?

27. The Panel found that continued reliance on ethinyl estradiol is appropriate, given that it is closely related to the natural ligand for the receptor of interest in this assay. The doses were specified but some panel members felt that the protocol evaluated by the Panel did not specify the methods for using this positive compound, nor were any methods found for the use of expected negative controls. Given that "the uterotrophic assay is likely to encounter both false positive and false negative events" (Phase 2 report, pg 45), before the assay is broadly implemented, some Panel members felt that it will be necessary to state in advance the acceptable ranges of response , and the acceptable rates of false positive and false negative results.

28. Whether such information on response ranges is available or necessary for all compounds or all doses is an issue for the PRP, and the information on false positive and negative results may be found in the contingency percentages provided. As importantly, a better understanding of the false positive rate for the Uterotrophic Bioassay is needed, and this could be determined by the evaluation of more true

negative compounds. The Panel recognized this as a key point which will be addressed in the next steps of the development of a Test Guideline and in the refinement of the protocol.

9) Are the strengths and/or limitations of the Uterotrophic Bioassay adequately accounted for and described in the protocols?

29. The Validation Management Group addressed the standardization of the performance of certain critical parts of this assay, and the Panel agreed with this as a necessary part of assuring similar performance at dissimilar sites. Procedures for ovariectomy and excision and *in vivo* animal handling are specified clearly. There has been some considerable evaluation of how different levels of dietary phytoestrogens can affect the outcome of the assay; no doubt other important dietary factors will become manifest as the assay is more broadly used. Other factors that should be specified include ascertaining that littermates are not put in the same treatment groups. The route of exposure, the doses used for specific chemicals, and the statistical methods used to analyze the data, could all be limitations if improperly performed.

10) Are there any editorial/technical corrections necessary for the proposed protocol?

30. In addition to the voluminous background literature supplied by the Sponsors, the Panel reviewed four study protocols generated to assess inter-laboratory variability. Because these protocols did not specify dose-setting, statistics (including thresholds for calling chemicals positive or negative, strong or weak), or acceptable ranges of change for specific known positive or negative control compounds , they are considered study-specific protocols. It should be possible to construct a final Test Guideline accounting for all of these. Small additional note: some members stated that the final protocol should clearly specify the range of acceptable days after ovariectomy when the animals may successfully be used in the assay. For example, Protocol C and CX in Phase-2 stated 14 days; this had been increased from 10 days in the prevalidation work in Phase-1.

11) Is the Uterotrophic Bioassay relatively insensitive to minor changes in protocol?

31. The influence of dietary phytoestrogens, bedding, housing conditions and (in a limited way) vehicle, was critically evaluated and (with the exception of the phytoestrogens) these were found to play a minor role in the performance of the assay. The final Test Guideline, when published, should contain a section that explicitly describes the acceptability boundaries of things like age (for immature animals the current protocols are very clear on this and they give a minimum for the OVX), and phytoestrogen content, so that there is no misunderstanding about what the “minor” and “major” variables are. One issue of concern for some Panel members was that so few animals are used. Some Panel members felt that the loss of only 1 animal from a group will materially reduce the statistical power of the assay at that dose level but this also depends on the CV and percentage of weight change. While it is recognized that animal number is determined after consideration of a large number of factors, the design of the final protocol should be such that a small loss of animals should not have a disproportionately large impact on the assay. While the Panel recognizes that this was taken into account in the study design, some members of the Panel still have some concerns with the small number of animals per group.

12) Are there any patent or proprietary issues that will inhibit the use of the Uterotrophic Bioassay?

32. The Panel was not aware of any patent or proprietary issues that would prohibit a uterotrophic bioassay from being presented for global use.

13) Are the apparent level of training and expertise required to conduct the Uterotrophic Bioassay reasonable for its wide use?

33. The Panel found that the key skills involved in performing the Uterotrophic Bioassay are those of basic animal handling, oral gavage, and prosection (i.e., necropsy dissection) techniques. While it may not be true that all the personnel in any professional lab possess all these skills, it is also true that these are not advanced techniques requiring intense or extensive training. Thus, it seems probable that any lab that maintains high quality of animal care and technical proficiency should have the skills, or the ability to acquire the skills, necessary to perform this assay.

14) Are the necessary equipment and supplies relatively easy to obtain?

34. The equipment and supplies are those of any professional toxicology lab, so their procurement should not be difficult. Maintaining the balances in good working order will be necessary, and the frequency of calibration should be specified, or the final protocol should specify that the assay is to be carried out under Good Laboratory Practices guidelines, which address the maintenance and calibration of lab equipment.

15) Is the method cost-effective, relative to the cost of conducting other *in vivo* assays?

35. The assay lasts only 3 days and weighs just a single tissue and thus will be inexpensive compared to any other in vivo test. Its cost-effectiveness stems from the fact that it evaluates a specific type of biological activity that is relevant to other mammals, including humans, by measuring a response in a complex living system. For the type of information gained, the assay's cost is low.

16) Is the time needed to conduct the Uterotrophic Bioassay reasonable?

36. By looking only at 3 days after the start of dosing, this assay is designed to find the relatively acute responses to oestrogens. This assay acts as a flag, an early warning, to identify a type of biological activity. The time required to obtain this information is quite short by any standard.

17) Has there been adequate consideration and appropriate incorporation of animal use, refinement, and reduction in the protocol, e.g., the group size of six animals?

37. The Panel took note of the explicit evaluation of group size, and agreed that n=6 appears to provide sufficient power for the purposes of the assay. Several panel members would have preferred a slightly larger group size (which would provide greater confidence in the answer, and would help protect the assay from the unanticipated loss of some animals), while on the other hand some other Panel members spoke against the unnecessary animal use involved in looking at both the juvenile and the OVX assays given that an initial assessment had shown that the two assays gave comparable results with the strong EE, not the weak agonists. Overall, n=6 appears to be a justifiable compromise. In terms of refinement, from an animal welfare perspective, it appears difficult to justify the continued advocacy of an animal model where surgery is needed (i.e. the OVX assay) when a non-surgical option (juvenile model) is available, that has been shown to provide comparable data and is widely used.

Test Method Data Quality and Sufficiency

18) Is there evidence that the data generated are of sufficient quality, including adherence to the protocol? This might include evidence that the data were or were not generated in compliance with Good Laboratory Practices.

38. There was a general agreement that the data presented were of sufficient quality, however, it was not clear whether compliance with GLP was required or merely recommended (Kanno et al. Environ Health Perpect 111:1530-49. 2003): "the laboratories were requested to perform these studies in compliance with OECD GLP and most, but not all, did so."

19) Are the data provided in sufficient detail to evaluate the results and performance of the Uterotrophic Bioassay for its proposed use? If not, what is specifically lacking?

39. The Panel generally agreed that the data were presented in sufficient detail to allow evaluation of the results. However, some concerns remained for some Panel members, as noted above: other steroids that have a possible uterotrophic effect (androgens, glucocorticoids, etc.) may have influenced the outcomes of the assay; only one positive anti-oestrogenic compound was tested and only one negative compound was used which may influence the assessment of the false-positive rate; and another limitation was that not every lab has tested all the doses of the weak agonist provided in the Phase 2-dose response study.

20) Were the characteristics of the test substances selected adequate to demonstrate the performance of the Uterotrophic Bioassay for its intended use as an *in vivo* screen for oestrogen agonist and antagonist activity?

40. The panelists agreed on the correct selection of compounds for the study, but the majority believed that the number was too low for antagonist detection. In Phase-1, only one antagonist was included. The Panel considered this an inadequate number of estrogen antagonist test compounds to evaluate the anti-oestrogenic predictive capacity of the Uterotrophic Bioassay. In addition, some Panel members expressed the opinion that one negative substance at only one dose was not enough to evaluate the specificity of the Uterotrophic bioassay. The Panel agreed that these areas need further exploration as the validation process moves forward to a Test Guideline.

21) Does the selection adequately represent the types of substances for which the test method is proposed to be used? Is it then appropriate to generalise the performance of the method for all test substances or are there important limitations on the applicability of the Uterotrophic Bioassay to certain test substance?

41. The concern of the members was not about the selected substances (all of them are known as weak agonists) but on the ability of the assay to detect potential hormonal activity of unknown chemicals. For this reason, most of the panelists agreed on the need for a tiered testing strategy with relevant *in vitro* methods, to classify a compound as an oestrogen-like substance.

42. In the case of Methoxychlor, with its enhanced activity due to metabolism, it was commented that this represents just one single example for this class of compounds. Most of the panel felt this was an appropriate first step, but was inadequate to conclude that the test, in isolation, will correctly identify those compounds that require metabolic activation to be either active or more active.

22) Was the use of the test substances in dose response experiments adequate to demonstrate the toxicological performance of the Uterotrophic Bioassay for its intended use as an *in vivo* screen for oestrogen agonist and antagonist activity? If not, why not?

43. The Panel agreed that the compounds that were tested were appropriate. One member stated that an important source of inter-laboratory variability was missed, since the doses of the weak agonists were not determined in an independent selection by each laboratory. The Panel noted that the number of antagonists was insufficient and that additional metabolically-enhanced positive compounds are desirable. The Panel noted the need for additional negative compounds to adequately demonstrate that the Uterotrophic Bioassay can routinely distinguish weak agonists from negative compounds.

23) Comment on the adequacy of the statistical/analytical methods used to evaluate the performance of the Uterotrophic Bioassay.

44. While the Panel acknowledged the extent of analysis that was performed, there was considerable discussion and debate around how to interpret the variability between laboratories, and which data analysis could have fitted better. It will be necessary to specify the statistical methods in an international Test Guideline.

24) Based on the Submission Package, are the results of the Uterotrophic Bioassay relevant and predictive for possible oestrogen agonists and antagonist?

45. The members stated that the information on the submission package suggests that the uterotrophic assay is a predictive tool for possible oestrogenic compounds, although its performance for identifying antagonists remains to be fully tested and defined.

25) Does the Submission Package adequately support the utility of the method for the regulatory use in hazard assessment of chemical substances that may have the potential to act as oestrogen agonist and /or antagonist? If not, why not?

46. There was a general agreement on the utility of uterotrophic bioassay to identify oestrogen agonistic compounds. However, sensitivity and specificity for the Uterotrophic Bioassay have not been sufficiently evaluated, in particular due to the limited number of negative test chemicals. The Panel agreed that the appropriate regulatory use of Uterotrophic Bioassay is as part of a larger group of tests, and not in isolation.

Determination of Test Method Reliability (Repeatability/Reproducibility)

26) Have the intra- and inter-laboratory reproducibility of the Uterotrophic Bioassay been adequately evaluated

47. A concern emerged about the intra-lab variability, which some Panel members felt was not tested sufficiently either for antagonists nor negative compounds. On the other hand, the dose selection, which is an important source of variability, was not addressed in the protocols reviewed by the panel, largely because what was reviewed were protocols for the performance of specific experiments.

27) Taking into account the objective of providing a test Guideline that can be used widely and internationally, comment on allowing the necessary flexibility in the selection of strain, diet, bedding, vehicle, and other conditions. Is there evidence to support that these differences significantly affect data quality (reproducibility, sensitivity, etc)?

48. The Panel noted the degree of statistical analysis that was performed on each of these variables, and most members agreed to the conclusion that most of these variables are not of critical importance. Some members thought that it is important, with regard to reproducibility of the test, to have information on technical skill and laboratory experience.

28) Was the reproducibility of the test adequately evaluated using coded (blinded) samples?

49. The evaluation of reproducibility for oestrogen agonists was done in an acceptable way. However, before an evaluation of reproducibility can be completed , more data needs to be generated for oestrogen

antagonists and negative compounds so that assay performance can be assessed for correctly identifying these substances.

29) Considering the variability inherent in all chemical and biological test method, are the results obtained with the Uterotrophic Bioassay sufficiently repeatable and reproducible?

50. The majority of the Panel agreed that the results obtained with the Uterotrophic Bioassay are sufficiently repeatable and reproducible. An observer pointed to the results of the Monte Carlo simulation presented to the PRP suggesting that the between-laboratory reproducibility is not good enough for the test to be considered reliable for distinguishing between weak agonists and non-agonist test substances.

Other Considerations

30) Considering the need to employ the Uterotrophic Bioassay internationally, can the test method be readily transferred among properly equipped and staffed laboratories? Specifically comment on the following:
a) Is the Uterotrophic Bioassay relatively insensitive to minor changes in protocol?
b) Are there complications or limitations that have not been addressed by the protocols?
c) Can the test method be readily transferred among laboratories properly equipped and staffed?

51. The opinions of panel members on these questions are summarised below. Consensus was not reached on all of these issues.

52. The PRP members suggested that the test method could be transferred if the following recommendations are taken into account:

- An acceptable range of phytoestrogen in the diets should be determined and maintained for the labs conducting this assay.

- An acceptable range of animal body weights should be specified.

- Restricted caloric intake should be considered.

- The importance of the 'competency' of the testing laboratories should be stressed to allow the Uterotrophic bioassay protocols to be transferable between laboratories.

- Guidance on how to interpret data, including the death of animals at high doses, or how to interpret a modest increase in uterine weight, should be provided. Such interpretation of data is pivotal to *in vivo* testing in general, but if there are specific data interpretation issues for the uterotrophic assay, these should be highlighted.

- Guidance for dose selection for chemicals of undetermined toxicity should be provided. This should include the criteria for setting the high dose, consistent with normal testing requirements to avoid unnecessary pain or mortality of the animals**.** Such dose-selection decisions are an integral component of study design for *in vivo* tests in general.

31) Is there any other information that should have been added to the Submission Package, published or un-published?

53. There are several additional items that some Panel members felt should have been included in the Submission Package. An example is the decisions for the selection of the test substances, although

selection criteria were noted in the reports. For example, one Panel member was concerned that there is evidence in literature for weak oestrogenic activity of DBP(1). The reference data cited in the package for the negative compound, DBP, is negative for the oestrogenic activity, but the comparison between *in vitro* results, the Uterotrophic Bioassay results, and the reference results were not clearly presented.

(1) Jobling et al. (1995) Environ Health Perpect 103:582-7. Harris et al. (1997) Environ Health Perpect 105:802-11. Zacharewski et al (1998) Toxicol Sci 46 :282-93. Tollefsen et al. (2002) Mar Environ Res, 54:697-701. Yu, et al. (2003) Wei Sheng Yan Jiu 32 :10-2.)

32) Has the uterotrophic bioassay been sufficiently evaluated and has its performance been satisfactorily characterized by the OECD validation program to support its proposed use for screening the potential of substances to act as oestrogen agonists and antagonists *in vivo*?

54. The Panel is confident in the robustness of uterine biology, and there is consensus that agonists will be identified by the protocol as described. There is broad consensus that no other *in vivo* or *in vitro* test for endocrine disruptors has been subject to this degree of evaluation. There is agreement that there is less confidence about the ability of the assay to identify antagonists. Due to the use of only one negative compound, there is limited confidence in the ability of the Uterotrophic Bioassay to correctly separate weak agonists from inactive compounds. The final protocol needs specifics for dose selection, the statistical approach that will be used, the criteria that will used to separate "good data" from "bad data", and the criteria that will be used to call a response positive or negative. In short, there are still a number of details to work out and specify before the assay is considered ready for general use.

33) Based on the information provided in the submission package, does this method adequately identify the potential for test substances to act *in vivo* as possible oestrogen agonists and antagonists?

55. The Panel agreed in the ability of the Uterotrophic Bioassay to identify strong agonists but the responsiveness to androgens is still unclear. Agreement on the data for identifying weak agonists was less complete, and having evaluated more compounds would have given greater opportunity for agreement, but the majority of the Panel agreed that the test was suitable for this purpose. However, some Panel members stated that clear criteria for data acceptance and interpretation are necessary, e.g. how to interpret a modest increase in uterus weight.

56. The Panel reached consensus that the test was not sufficiently evaluated in regards to its ability to correctly predict negative compounds, or those requiring metabolic enhancement. Additionally, more specification is needed for setting doses, using the proper statistical tests, defining data quality criteria, and the final prediction model, before the assay can be said to be validated as fully capable of identifying strong and weak agonists, or antagonists. There was not consensus on the issue of having a prediction model. There was general consensus that this test should only be used as part of a tiered testing approach as specified in the Conceptual Framework of the OECD for the testing of endocrine disruptors.

57. Some members of the Panel did not agree that the test has "been validated for widespread use". Other Panel members agreed that the underlying biology will allow the identification of the strongest oestrogen mimics. One observer stated that the work to date represents a good start at a pre-validation, but nothing more. The entire Panel recognizes that more work needs to be done to better understand the limits of the test and to make it truly ready for global deployment.