Responses from Peer Review Panel members to Charge Questions.

20th October, 2003 – First Set of Peer Review Panel Responses

This section records all comments from the Panel Members and Observers for the Uterotrophic Validation Peer Review on the first round of charge questions. These were discussed at the teleconference on Monday, 20th October, 2003.

The Test System
Is the choice of the rodent uterus, specifically in this case the rat uterus:
a. Biologically relevant for the detection of oestrogen agonists and antagonists in vivo?
Member 1 The rodent (rat) uterus IS biologically relevant: it is endowed with significant numbers of alpha ER's, and activation of those receptors elicits numerous molecular and biochemical changes that ultimately result in a significant increase in cell number and fluid content, which can readily be measured on a balance. That these effects are mediated via the ERalpha is indicated by the attenuation of these responses in ERKO animals.
Member 2 Yes, the choice of the rodent uterus is appropriate. It should be recognized, however, that there are significant differences among rodents (rats, mice and to a lesser extent, voles have been the primary species studied.) Also, significant differences in the sensitivity of the uterus to estrogens exist among strains within a species.
Member 3 The Uterotrohic bioassay is considered biological relevant for the detection of oestrogen agonists/antagonists since the biological basis for the Uterotrophic bioassay is the role of oestrogens in the natural oestrus cycle. The bioassay is a robust and rapid *in vivo* screening for possible oestrogen agonists/antagonists, based on the response in oestrogen sensitive tissue. In the uterus ERα is present at significant levels whereas ERβ is present at moderate levels which limits the detection of ERβ agonist/ antagonist with the Uterotrophic bioassay. The bioassay is biologically relevant for the detection of agonists/antagonists, however, what does an increase (small) in uterine weight mean biologically, is it an adverse health effect? The assay seems quite insensitive compared to mammary gland dysgenesis in adult mice exposed to xenoestrogens in foetal life. Bisphenol A induced mammary gland dysgenesis at concentrations 4000 fold lower than those required to induce an uterotrophic effect. There may also be a risk for false negative results because of the low sensitivity of this test. This could possible be solved by measuring the expression of one or two oestrogen regulated genes or proteins.
Member 4 The rodent/rat uterus response in animals with low endogenous oestrogen levels is biologically relevant for the detection of oestrogen agonists; modifications of the assay providing (endogenous or exogenous) oestrogen along with a test agent are suitable for the detection of oestrogen antagonists.
Member 5 Yes, rat uterus growth is a relevant biological phenomenon for detection of chemicals with estrogen-like biological effect
Member 6 The rat as well as mouse uterus are of sufficient biological similarity to the human uterus such that oestrogen agonists and antagonist can be detected. It, however remains to be proven whether the sensitivity of detection is sufficient to allow identification of weak agonists and especially antagonists. This in view of the fact that some chemicals have been detected with different sensitivities within rodent species and between different rat strains. Therefore one has to expect even more pronounced interspecies variations between rats and humans.

Responses for 20th October, 2003, teleconference

Is the choice of the rodent uterus, specifically in this case the rat uterus:
a. Biologically relevant for the detection of oestrogen agonists and antagonists in vivo?
Member 7 The existing literature indicates that the rat uterus is a biologically relevant test system for detecting oestrogen agonists and antagonists. It also responds to non-oestrogens, e.g. androgens.
Member 8 Yes.
Observer 1 Comment: The test is relevant for detection of chemicals that produce uterotrophic effects in rats. The assay *alone* is insufficient to specifically detect estrogen agonists. As noted in chapter 188 of the BRD non-estrogens can induce increases in uterus weight. If this statement is true, then additional studies would be needed to confirm that a uterotrophic effect is estrogenic.
Panel Observer #3

Responses for 20th October, 2003, teleconference

b. Mechanistically adequate and sensitive for the detection of oestrogen agonists and antagonists in vivo?

Member 1

The response of a target tissue to estrogen is complex, and has been broken down into many component responses at the biochemical and molecular level. For the purposes of toxicity testing, we are less concerned with how a compound works, and more concerned whether the body "sees" the compound as an estrogen. This test IS mechanistically adequate because of the reasons noted above (it has estrogen receptors and they elicit numerous responses). Based on the dose-response patterns presented in these documents, the assay is variably sensitive: it appears more sensitive to some compounds than to others. It certainly is sensitive to strong agonists such as 17β estradiol and diethylstilbestrol.

Member 2

Yes, sufficient mechanistic information is available to test for estrogens and antagonists with the proviso stated above about species and strain differences. One area that could require more mechanistic investigation is the cellular mechanisms of water imbibition in the stimulated uterus. A great deal of work has been done on the mechanisms of cellular hyperplasia but less on water transfer and, since one measure used in the uterotrophic assay includes the weight of the water, this would seem worthy of interest.

Member 3

The measurement of luminal epithelia cell height is for some weak antagonists e.g. Bisphenol A reported to be more sensitive than increases in uterine weight measured by the Utreotrophic bioassay.

Measurement of uterine cell proliferation by BrDu labelling or PCNA, or the expression of oestrogen regulated genes or proteins are other sensitive endpoints for the detection of oestrogen activity.

Member 4

The test system is more than adequately characterized on a mechanistic basis.

Additional biochemical/molecular markers and histological changes can *assist* in an evaluation of agents, although at present none of them has proven to be consistently more sensitive than the uterine weight response when applied to range of chemicals or assessed with different routes of application.

Member 5

Yes.

Member 6

The test is based on a single endpoint, the increase of uterus weight after stimulation with hormonal active chemicals. Other oestrogen responsive tissues (Detailed background review document; page 24) with different metabolism and without rapid turnover of cell growth have not been taken into account.

Apparently the immature rat version has been the first test that has been developed. It is not understandable why the assay has been "improved" by using OVX animals. AT least, if possible the equivalence of both assays should be stated for animal welfare reasons. The argument that not all laboratories have access to animals with a defined birth date is very weak and should not be taken into account due to animal welfare reasons. Even though on page 65 it is remarked that in most countries animals with a requested age can be supplied. The study showed that either of these two basic protocols can be used, although it should be noted that this conclusion is based on results from the widely differing numbers of laboratories that used the four protocols. The data suggest that the need for two protocols is scientifically not necessary, and that the protocol based on immature animals should be sufficient for the most purposes. It is most important the such scientifically unjustified flexibility is not allowed to become incorporated into any OECD test guideline.

Furthermore, the quality of ovariectomy is an essential component in the analysis of effects. (see Crit. Rev. Tox Owens and Ashby, 2002; page 474 and Zackarewski et al, 1998).

Responses for 20th October, 2003, teleconference

b. Mechanistically adequate and sensitive for the detection of oestrogen agonists and antagonists in vivo?
Member 7 The mechanistic basis of the test system response has been elucidated and is reasonably well documented. Evidence suggests that it is sufficiently sensitive to distinguish weak oestrogen agonists.
Member 8 Yes.
Observer 1 Comment: The test is mechanistically adequate for the detection of chemicals that cause uterotrophic effects in the rat. As noted above, it is apparent that additional study would be required to confirm that the effects of uterotrophic chemicals are estrogenic (see above).

Responses for 20th October, 2003, teleconference

c. Toxicological an appropriate choice for the detection of oestrogen agonists and antagonists *in vivo*?

Member 1

Practicality and cost would suggest that one would want the smallest species with a consistently estrogen-responsive uterus. A limited number of rat strains are used for the vast majority of reproductive toxicology testing. In contrast, mice are the genetic workhorse of science, and have been bred into dozens or hundreds of different strains. It is the experience of this Member that there is more inter-strain variation in mice than in rats, although the data summarized in the background documents seems to suggest that there is no significant difference between the mouse strains tested. This panelist is comfortable using rats for this test, and would prefer them to mice (too variable) or rabbits (unnecessarily large and expensive).

Member 2

Yes, at present an animal uterotrophic assay is clearly the best assay for putative estrogen agonists and antagonists to assess information for human hazard assessment. In my opinion, in-vitro assays currently available may provide some screening information but do not provide adequate data for assessment of animal or human risk assessment. Again, the species (laboratory rat or mouse) and the strain should be considered. I was impressed that there was considerable concordance among the test laboratories in the responses even though some different strains (and sub-strains)were used. This would suggest that the responses across strains are sufficiently robust. However, I think more research is needed on strain differences if a single strain is to be recommended. There are published data using DES in mice suggesting very low doses (in the microgram/kg. range) can stimulate the uterus.

Member 3

The rat is considered the most appropriate species since the rat is usually used in developmental and multi-generation studies, and in mice a higher level of differences in responsiveness has been reported.

In the phase-one and phase two studies different rat strain were used, mostly Sprague-Dawley and Wistar rats, however, no evidence of strain difference was reported. However, difference in rat strain sensitive is reported for Bisphenol A where Fisher 344 rats and Alpk:AP rats were more sensitive than the Sprague-Dawley rats. This difference was explained to be associated with differences in intermediate or late gene expression of oestrogen regulated genes, which indicates the need to also include the measurement of the expression of oestrogen regulated genes.

Member 4

In principle, both rats and mice are an appropriate choice:, Mice, due to a lower body weight, may be useful in cases of limited amounts of test compound. But, in light of the available data, the rat appears to be a better choice because of lesser (no significant) differences in the responsiveness among strains.

Member 5

Rat is a commonly used animal in toxicological testing, housing, diet etc. are clear, bigger than mouse and work is easier with it**.**

Member 6

On page 87-89 of the background review document the authors describe a positive uterotrophic result can also occur with non-oestrogens e.g. androgens, progestins growth factors. In order to decrease the level of false positives the authors suggest performing precursors assays such as ER binding. It is questionable whether an Uterotrophic test is still necessary if the chemicals have demonstrated positive ER binding activities and a response in transcriptional/reporter gene assays. A positive in vitro result should not automatically lead to an additional in vivo experiment. Since the authors stress several times that the uterotrophic test should be used in a tiered test strategy in addition to in vitro tests the added value of this test is limited (receptor binding tests can be combined with biokinetic modelling).

The Uterotrophic bioassay aims to detect also weak oestrogens and antioestrogens. Since very high concentrations (up to 1000mg/kg) have to be applied in order to receive some biological responses it is questionable if these results are relevant for human hazard assessment. For some chemicals and protocols the selected dose was already lethal (mainly protocol A but also cases in protocol B and C). Since a prediction model is missing in the protocols, it cannot be judged which increase in uterine weight should be considered as biologically significant Without the use of more negative control chemicals it is impossible to fully assess the predictivity of the assay in terms of its specificity and its sensitivity. An additional study to analyse the reliability and relevance of the assay for detecting anti-oestrogens is necessary to evaluate the predictive power for this purpose.

Member 7

Yes, and is probably preferable to either the mouse or the rabbit for logistical reasons.

c. Toxicological an appropriate choice for the detection of oestrogen agonists and antagonists *in vivo*?
Member 8 Yes, rodents are optimal in view of large litters and historic information, mouse and rat seem equally sensitive, rat is somewhat larger so easier to work with. Strain differences are small and have no major impact on the results.
Observer 1 Comment: The test is appropriate for detection of chemicals that are uterotrophic in rats (see above). Additional data would be required to determine whether an observed uterotrophic effect is estrogenic.

Responses for 20th October, 2003, teleconference

d. Consistent with the use of animals to obtain *in vivo* hazard information for human hazard assessment?
Member 1 This test should tell us if a compound poses a hazard to a living system. Effects in molluscs or fish can be directly addressed but effects in humans must be measured by proxy; it is appropriate to use a living intact mammal as a proxy for human effects. Humans are not naked ambulatory receptors, but complex systems engaged in active homeostasis. It is important for the test to contain as many of these homeostatic processes as possible.
Member 2 Yes, the use of animal models is essential for the detection and verification of putative estrogenic mimics or blockers.
Member 3 The Uterotrophic bioassay is an *in vivo* screening assay for oestrogen agonists and antagonists. The assay is based on the assumption that a statistically significant increase/decrease in uterine weight implies that there has been an exogenous source of oestrogenic/antioestrogenic exposure, which in itself is regarded as a toxicological concern, because it is then considered to affect the endocrine system in humans. However, the Uterotrophic bioassay results will need careful examination before triggering significant toxicological testing. This indicates a clear need to establish a defined process to accept data and to consider the weight of evidence from all *in vitro* and *in vivo* data before concluding whether the bioassay results are valid. Positive results in the Uterotrophic bioassay, a mechanistic screening assay, is considered not be used alone to define endocrine disruption via an oestrogen mode of action. Classification and labelling is considered not to be based on *in vitro* and *in vivo* screens e.g., transcriptional assays or the Uterotrophic bioassay, but await the results of more definitive testing protocols for adverse health effects such as multigeneration or developmental studies. In the phase 2 study it was shown that the increase in uterine weight following exposure to the phytoestrogen Genistein was higher when Genistein was given by the subcutaneous route compared to the oral route. However, in this case the strongest response may not always be the most relevant response for human risk assessment.
Member 4 The use of animals in hazard characterization is warranted in light of the important role of toxicokinetics in determining biological activity (e.g. metabolizing pro-estrogens to estrogens; route of administration and/or bioaccumulation affecting judgements on relative potency). Moreover, in vivo assays are indispensable for the detection of endocrine active chemicals operating by indirect modes of actions that are not yet detected in *in vitro* assays or by *in silico* (QSAR) approaches.
Member 5 Yes
Member 6 It needs to be clarified why some chemicals that are positive in the uterotrophic assay are negative in the classical tests such as multigeneration studies and dev. Tox. tests. Are they really false positive? Since only one negative substance has been included in the study it is difficult to define the predictive value of the tests. The use of animals with little endogenous estrogen prevents an evaluation of effects of weak estrogens and antiestrogens during an intact hormonal cycle. Synergistic or additive effects will not be detected. (E.g. critical review tamoxifen 451). As discussed before since the prediction model is missing the relationship between ill-health and uterine weight increase is unknown.
Member 7 Yes, its use is commensurate with traditional laboratory animal data inputs (e.g. NOAEL, LOEL) to toxicological risk assessment.
Member 8 Yes, although the relevance of subcutaneous versus oral dosing is an important issue. For hazard, both are relevant in their own way.
Observer 1 [Secretariat: No reply received from panel observer on this point.]

Responses for 20th October, 2003, teleconference

Test Method and Protocol Description
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?
Member 1 The test and the pertinent variables are described with admirable thoroughness. My only suggestion is that the protocol itself specify the number of days that should elapse between ovariectomy and treatment. This is addressed in the detailed background review document, but I could find no mention of it in the protocols themselves. Also, I would suggest that the batch of diet be recorded. This was mentioned in the supporting text of the protocols, but not listed as a specific endpoint, and I believe it should be.
Member 2 The protocols are described in adequate detail but I would add one item. I think the corn oil used as the vehicle should be tocopherol stripped rather than store-bought corn oil. Corn oil can contain estrogenic substances and the commercial oil available can be variable. I believe the USEPA uses only stripped oil as a vehicle. This is particularly relevant since corn oil was used in 12 of the 19 test laboratories.
Member 3 Overall the test method and protocol is described in sufficient detail, I only have some remarks: The immature female rat protocol should include that pups outside a 35 to 50 g bodyweight should not be used on pnd 19 (start of exposure) since pups smaller than 35 g may not respond as well as pups larger than 35 g whereas pups larger than 50 g may begin to secret E_2 by day three of the assay when they reach or exceed a body weight of 60 g. In Annex 3 phase 1 study, and Annex 2 phase 2 study, under the section *results for individual animals* it is described that organ weights at necropsy should be measured, however, which organ weights, are the measurements related to systemic toxicity of the test chemical. The antagonist ZM was used together with EE to test for antioestrogenicity, however, maybe a group only receiving ZM alone should have been included so that information regarding effects of exposure to ZM alone were available. Consider the possibility that complimentary techniques are needed to confirm or refute equivocal results. The most promising alternative as described in the critical review document may be the development of gene assays to capture the profile of a subset of genes that are specifically up or down regulated by oestrogens. At the Eurotox 2003 congress Orphanides G held a very good lecture with the title; *How do natural and synthetic estrogens induce cell proliferation and differentiation? A transcriptional view of the uterotrophic response*.
Member 4 The model protocols provide sufficient details how to conduct the assay, along with information on variations of protocols and how to evaluate the outcome.
Member 5 Described correctly, sufficiently. All protocols are clear and can be reproduced with some practice. Endpoints are relevant, variations are clearly shown.
Member 6 The protocols in Annexes of the phase 1 and 2 report appear in some cases as a mixture of a study plan and a test protocol. e.g. Phase 2; Annex 2; pp 94 …..Characterisation of the test substance is the responsibility of the chemical supplier and those managing the chemical repository. It is not the responsibility of the lead or participating laboratory….page 104 The heading “statistical analysis of the results” is only describing that the data have been delivered to the leading laboratory. There is no guidance how to interpret the results. The description of the different versions of protocols in chapter 3 and 4 is a summary of possible variables of the test. These variables are sufficiently described. However, since the protocol is broken down in “building blocks” it can be combined as wanted by the future users. This can lead to protocol variations, which are not validated and the reliability of this tests has not been proofed.

Responses for 20th October, 2003, teleconference

Test Method and Protocol Description
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?
Member 7 The protocols have been written in "test guideline form" rather than as SOPs (which typically provide more detailed technical information and are more usual in validation studies). It is noted that a video of the animal procedures was available to support the protocol and facilitate standardisation between the participating laboratories. However, it is also evident that some laboratories did not conduct the assays as intended. Protocols are not the final versions; these need to be refined following the outcome of phase II, incorporating the various learnings re maximum phytoestrogen content of the diet, solubilisation of test substance, selection of appropriate doses based on prior conduct of range-finding studies, etc. Sections on data analysis and data interpretation need to be incorporated. Known limitations of the assay need to be incorporated.
Member 8 Yes.
Observer 1 Comment: No. The protocols described in Annex 1 Phase 1 and Annex 2 Phase 2 Reports are early developmental protocols "designed to enable variation between laboratories to be investigated and protocol refinement to be proposed". A key part missing from all protocols and the discussion in Chapters 3 and 4 of the BRD is the definition of procedure laboratories should follow to establish the starting test substance dose and dose intervals. The protocols used in this study specify the dosing regime to be used for each chemical. Since variation in dosing between laboratories can cause significant variation in predictions of toxicity this validation study did not assess the capacity of the uterotrophic assay to correctly predict the uterotrophic effects under real-world conditions where the dosing regime is not specified.

Responses for 20th October, 2003, teleconference

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.
Member 1 It's a little late to be asking this question, isn't it?… if you've done the tests already, then why bother with this? But the answer to this question is Yes, there is adequate detail for the material and equipment required to conduct the test, with the exception that the length of time between OVX and treatment should be specified in the adult OVX protocol.
Member 2 Yes, but see above. [Secretariat: presumably the comment on vehicle in the previous question]
Member 3 In my view it is good described, however, I have not had the opportunity to do the Uterotrophic bioassay in the lab, so my practical knowledge is limited.
Member 4 The model protocols are complete and adequate in detail.
Member 5 Description is adequate, especially because no specific equipment is necessary. In the protocol I did not find the specification of measuring, I mean body weight with 1 g precision and uterus weight 0.1 mg (but these are given in EHP series).
Member 6 Chapter 3 and 4 of the detailed background review of the Uterotrophic bioassay provide a detailed analysis of protocol variables that have been described in the literature. This can be confusing since the document is also citing versions of the Uterotrophic tests that have not been validated. It should be stressed that only the test protocols described in the annexes of phase 1 report and phase 2 report have been validated. An integration or exchange of parameters such as the administration of test material via drinking water etc. will lead to a variation of the presently proposed version of the Uterotrophic assay. The protocols are not described as detailed as in a standard operation procedure, which can lead to misinterpretations. E.g. the necessary housing and husbandry conditions to increase the sensititivity and efficiency of the assay need to be clarified. The protocol requires personnel with consistently good technical skills, e.g. animal husbandry, dose preparation, and dissection. A specific training in animal experimentations such administration of test chemicals and killing of animals should be mandatory. In addition, the weight of the uterus used as toxicological endpoint is highly dependent on the isolation procedure. A high intra and inter laboratory variability has to be expected which can be crucial during the classification of weak oestrogenic compounds. Indeed, the crit review of the uterotrophic assay by Owens and Ashby (2002) clearly demonstrate that the dose response of 17β-estradiol and uterine weight (Table 4 , page 461) is not a very steep one (dose range 100000 vs relative weight increase 1.05-1.63), clearly demonstrating that weak agonist will be difficult to detect in view of a potentially high uterine weight variability resulting from handling procedures. Some protocols require the ovariectomy in order to avoid endogenous oestrogen. This operation is another crucial parameter of the test. Rests of the ovary can produce oestrogens that are interfering with the test (see comments above). A video is not sufficient for an untrained laboratory.
Member 7 Yes.
Member 8 Yes.
Observer 1 Comment: Description is adequate.

b. Description of what is measured and how the data are used to identify positive and negative results?

Member 1

Yes, the protocol is clear about what is to be measured and how the data are to be handled. It is apparently silent on how a positive result is identified (i.e., are you happy with simple statistically significant difference from control? And do you want the lowest effective dose? How are those data interpreted and handled??)

Member 2

Yes, the descriptions are adequate.

Member 3

A clear description regarding a positive effect e.g a 40 % increase in uterine weight should have been discussed in more detail, and when should historical control data be included in the interpretation of positive/negative results.

Due to the appearance of false positive and false negative with the Uterotrophic bioassay, the need for clear criteria for data acceptance is necessary e.g. maxima for acceptable vehicle control uterine weights, and how to interpret a modest increase in uterine weight (20 to 40 %).

Non monotonic responses (U-shaped response curves) are reported for oestrogen like chemicals, how to handle this should have been addressed.

I miss in the protocol a definition of the wet uterine weight (defined as including intraluminal fluid) and the blotted uterine weight (defined as having the intraluminal fluid removed).

Member 4

The description (and figures) provide sufficient details on what is measured and how to evaluate the outcome.

Member 5

All clear, it is easy to perform and evaluate from the description.

Member 6

A guidance how to interpret the results is missing in the protocols. It is not clear when a chemical has no effect and when a chemical has to be classified as active. (At least the paragraph on statistical analysis of the crit.rev.tox publication should be included into the protocol.) More detailed information will be provided by ECVAM's statistical review. There is no prediction model included that allows to judge if the uterine increase is scientifically relevant or not.

There is no indication in the test protocol for chemical testing (phase 2 report; annex 2) how to select test concentrations for unknown chemicals. On page 98 (phase 2 report; annex 2) it is stated ….all substances will be administered at pre-determined dose levels and will be coded to create a blind study. This is again part of a study plan and not guidance how to select concentration levels of unknown chemicals.

Member 7

The endpoints and data to be recorded are described adequately. In final, refined, versions of protocols submitted for formal validation the statistical analysis of the results and the manner in which the data are to be interpreted ("prediction model / data interpretation procedure") should be documented. This is not the case in these protocols.

Member 8

Yes.

Observer 1

Comment: No. An unambiguous statement of the prediction model is missing from all protocols. The description of endpoints to be measured is given.

Responses for 20th October, 2003, teleconference

c. Are there appropriate provisions for the use of reference control chemicals?
Member 1 Yes, ethinyl estradiol is specified as the reference control chemical, which is closely analogous to the natural ligand in situ. This is an appropriate choice, and it readily available in relatively pure form.
Member 2 Yes. My only suggestion is that the group might consider DES as a reference estrogenic substance since so much work has been done on this compound.
Member 3 Limited data is given on the antioestrogen ZM 189.154, information regarding chemical structure would have been helpful. In the Phase 2: Coded single dose study the phthalate DBP was included as a negative chemical, however, an increase in the uterine weight was reported for this substance, therefore, maybe more than one negative chemical should have been included in this study. In the phase 1 study a vehicle control group as well as an untreated control group is included, however in phase 2 only a vehicle control group is included. An untreated control group should also have been included since the vehicle may contain some phytoestrogens, and the animals may be stressed by the treatment which may have some impact on the results.
Member 4 The provisions for reference compounds (agonist and antagonist) are appropriate.
Member 5 Yes.
Member 6 There is no indication in the phase 2 annex 2 report how the positive reference control should be used. The results of the phase 1 report should be included. The protocol of phase 2 (protocol for testing of unknown chemicals) should include the indications how to use the reference control (storage, concentrations etc). There is no reference for a negative control included. Indeed, only one non-active chemical has been tested during the test evaluation! The protocol should provide a clear statement when a test can be judged as valid. This should be dependent on the results of the control chemicals (negative and positive) whereby the body weight, as an indicator for toxicological effects, must be taken into account.
Member 7 Vehicle and positive (ethinyl oestradiol) controls are included and detailed in the protocol (phase 2, protocol C). The final protocol should include a negative control (non-oestrogen agonist or antagonist), and details of other "reference chemicals" to be tested in the assay so that the ability of a laboratory to perform the assay correctly can be assessed. The results expected (means ± SD, ranges) need to be included as "assay performance criteria". Typically, these would be defined before a validation study commences. Experiments where the data for the controls and any other reference chemicals were outside the ranges deemed acceptable would be disqualified from the subsequent analyses.
Member 8 Yes.
Observer 1 Comment: The protocols do not provide guidance on the use of control chemicals.

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

Member 1

Such as they are, yes.

Member 2

Yes

Member 3

In the protocol, Annex 3 phase 1 and Annex 2 phase 2 I do not find any description of the strength and/or limitation of the Uterotriphic bioassay. It is only described that "If chemical causing and increase in the weight of the uterus thereby indicates that it has activity consistent with natural oestrogens. This description is considered too limited. However, in the Background review document the strength/limitation is described in more detail related to the immature versus ovariectomised rats, route of administration, strain differences, rats versus mice, etc.

The difference in gene expression in young immature rats versus adult ovariectomised rats, and how this could influence the results should have been discussed, since younger rats are shown to be more sensitive to oestrogens/androgens in early life.

One comment related to the use of immature rats is what about a litter effect on the outcome of the uterotrophic bioassay. It should be mentioned that littermates should not be assigned to the same group, or to include the possibility of litter effects as part of the statistical analysis. One other small technically related comment is to the use of immature female rats is that when using immature rats small weight changes in measured in small animals.

In the phase 2 study weak oestrogen antagonists were tested in the Uterotrophic bioassay, and they all showed an increase in uterine weight at one or more of the doses tested. However, other weak partial agonist may only increase the expression of early oestrogen regulated genes, which are not involved in cell division, and consequently, no significant increase in uterine weight will be found, although the weak agonist upregulate early oestrogen regulated genes.

Have to define performance criteria for data acceptance for uterotrophic results, such as a maximum uterine control weight of 45 – 50 mg in the immature rat and 16 – 18 mg in the immature mouse.

Member 4

The protocols and particularly the Background review document compile useful information on various versions of the bioassay and procedural variables. Other than in the CRC-(printed) version, the cited literature now appears to be complete.

With regard to protocol variables, I'm in favor of the recommendation (in Annex 2 of the Phase 2 report) to store samples of the rodent diet for an analysis (if necessary in the case of questionable results of the bioassay).

Member 5

Yes.

Responses for 20th October, 2003, teleconference

Member 6

The authors suggests to develop a tiered test strategy including in vitro and short term in vivo screening methods such as the Uterotrophic bioassay for the detection of endocrine disrupters. However, it is not sufficiently explained why a test chemical that has been shown to have a high affinity in a receptor binding test or a positive result in a transcriptional test should be tested in the Uterotrophic test.

The route of administration seems to be another variable part in the protocol (different responses to selected chemicals). Intravenous injection, intraperitoneal injection, subcutaneous injection, intramuscular injection, oral gavage, inclusion in the drinking water, food and dermal application have been described in the literature. On page 61 of the detailed background review document it is stated that the route of administration of the test chemical should be selected according to regulatory policy needs. However, only subcutaneous injections and oral gavage have been tested in the validation study. Only these administration routes are thus available for regulatory safety testing. Since it has been demonstrated that the route of administration is a sensitive part in the protocol the validation study should have defined the route of administration before stating for which specific purpose a test will be validated. A clear guidance has to be included into the protocol, which defines the route of administration for certain applications or chemical classes. In addition, an unacceptable high amount of dead immature animals have been observed after oral gavage administration of some test chemicals. The combination of oral gavage and immature animals should be excluded. It has not been explained how to evaluate dead or morbid animals in the statistical analysis.

The Uterotrophic Bioassay aims to detect weak endocrine active compounds. In the critical review document on page 484 it is written that there is sufficient variability so that some rate of false neg could occur with very weak agonist. Even with a stricter standardization some variability amongst laboratories and technicians may be expected.

However, the detection of weak estrogens will depend how the assay is performed (see comments above). Discussions if a chemical should be classified as weak estrogen or not are foreseeable. The suggested solution just to administer a higher concentration of the test chemical seems not to be a scientific solution since a very high test concentration such as 1000mg/kg is coming close to a toxicological dose and will not reflect an environmental exposure.

Since the tests aims to detect weak oestrogens e.g. environmental contaminants, the sensitivity and specificity is relevant for this class of chemicals. It is highly questionable if this bioassay can deliver the required information.

Member 7

Limitations and potential difficulties require better definition.

Member 8

Yes.

Observer 1

Comment: The fact that these are pre-validation protocols is clearly and appropriately stated.

Responses for 20th October, 2003, teleconference

e. Are there editorial/technical corrections necessary for the proposed protocol?
Member 1 Under "Reporting Requirements", the "Test animals" section should contain a statement of how long the animals should be ovariectomized before use, and should also state if there is a time that is too long (e.g., "should be ovariectomized for at least 14 days and no longer than 56 days"). Under "Test Conditions",a record should be kept of the batch of diet used (this is mentioned earlier in each protocol, but not listed in the "Reporting Requirements" section)
Member 2 Probably so, but that depends on what the group decides concerning changes in the protocol (e.g., comments about strain differences, use of DES as a test compound, and suggestions by other members of the Panel). The final editorial comments should be done later.
Member 3 In the protocols for ovariectomised rats it should be described that a period of 14 days for uterine regression should be included to obtain adequate responsiveness, and that the ovariectomy has to be monitored to ensure that it is complete, for example by vaginal smears or observation of the ovarian tissue remnants at necropsy.
Member 4 Aside from typing errors, there is no need for major corrections.
Member 5 No.
Member 6 The protocols in the annexes of phase 1 and phase 2 report has to be reviewed and parts which are belonging to study plans have to be rewritten
Member 7 Versions included with the phase 1 report are not definitive. Similarly, the example protocol in the phase 2 report cannot be considered adequate as a final protocol for use in a formal validation study (see omissions outlined in previous sections), let alone sufficient as the basis for developing a new test guideline.
Member 8 No major problems. I wonder however whether a preference for either the immature or the ovariectomized version could be given, even though the assay data show comparable results for both. Technical arguments (small young animals versus larger adult with neccessary ovariectomy) and animal welfare (ovariectomy) and cost issues have likely been discussed at some stage in the validation program. As regards dosing, oral versus sc dosing may give different results in young vs ovariectomized adults in view of differences in ADME. How (if at all) does the compound to be tested influence the protocol (age and route) choice?
Observer 1 Comment: Yes. As noted above the protocols lack a description of procedures to be used for defining starting test substance dose and dosing intervals. The prediction model or other data interpretation procedure is missing from all protocols. Both needed in order to complete a useful protocol a/o guidelines to be used generally.

Responses for 20th October, 2003, teleconference

Other Considerations
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?
Member 1 This was actually found in the entire Phase 2 report, which does a great job of examining the effects of procedural variation. Phase 2 does a remarkably thorough job of evaluating the responses from multiple labs, and exploring the causes of variation. It is clear that a minimal number of animals is used, such that any animal mortality more than 1 animal/group measurably reduces the sensitivity of the test. Other than that, there is consistency across laboratories and countries in their ability to identify compounds that stimulate modest increases in uterine weight. If one doesn't place great weight on the absolute values, this consistent response is quite remarkable.
Member 2 From the material presented it seems that the uterotrophic bioassay is relatively robust and can be used internationally as long as the protocol is followed. I would assume that publication of reports using the assay to test for estrogenic substances would be in peer reviewed journals and that only such information be used to make final decisions on risks related to putative estrogenic substances.
Member 3 After going through the results from the phase 1 and phase 2 studies the Uterotropic bioassay is considered relatively insensitive to minor changes in experimental conditions e.g. changes in animal strains, housing conditions, bedding and vehicle used. Considering the influence of diet it was summarised in the Phase 2: Dietary study that diets containing less than 325 –350 µg/g Total Genistein Equivalents (TGE) did not impair the responsiveness of the bioassay. However, a study by Ashby, 2000 (APMIS 108: 805-813) indicated that undefined components (not phytoestrogens) of the diets are eliciting estrogenic effects, of different magnitudes, via a common action on the pituitary gland.
Member 4 There is good reason to assume that the test method is rather insensitive to minor changes in protocol and thus can be readily transferred internationally.
Member 5 Relatively insensitive (means not very much sensitive) in respect of the same protocol in the outcome of positivity sensitive regarding the different protocols.
Member 6 The term "minor changes" should be more specified. Apparently the test is quite resistance with regard to diet, strains etc. However, it is very important not to change crucial parameters of the tests such as the age of rats etc. This should be more specified in the protocol/SOP.
Member 7 It would appear to be robust and relatively insensitive to minor changes, although it is clear that some laboratories had difficulties in following the protocols as intended.
Member 8 [Secretariat: No reply received from panel member on this point]
Observer 1 [Secretariat: No reply received from panel observer on this point]

Responses for 20th October, 2003, teleconference

b. Are there any patent or proprietary issues that will inhibit the use of the Uterotrophic Bioassay?
Member 1 No, which is one of the attractions.
Member 2 Not to my knowledge.
Member 3 I can not find any patent or proprietary issues related to the use of the Uterotrophic bioassay. However, maybe, food to reduce background level is one.
Member 4 There is no reason to assume that any patent or proprietary issues would inhibit the use of the Uterotrophic Bioassay.
Member 5 I did not find and do not know about.
Member 6 [Secretariat: No reply received from panel member on this point.]
Member 7 Do not appear to be any such issues.
Member 8 [Secretariat: No reply received from panel member on this point]
Observer 1 [Secretariat: No reply received from panel observer on this point.]

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

Member 1

Another attraction of this test is the large degree of response of the tissue, which contributes to its ease of performance. Any secondary school student could do the dissections. The degree of response means that reliable data can be obtained using a triple-beam balance in relatively primitive circumstances.

Member 2

The only part of the expertise required that worries me is the dissection technique. Standardizing where the uterus ends and the cervix begins can be a problem as well as not losing some fluid when the uterine fat is removed. The data presented suggest that an uterotrophic effect is found whether the measurement is wet-weight or blotted uterus so the latter might not be too big a problem. However, as this assay becomes widely employed it may become an issue. I think the drawings help and provide adequate instruction.

Member 3

I have not worked with the Uterotrophic bioassay in the lab, however, with my experience form working with experimental animals in toxicological testing I would consider that that the Uterotrophic bioassay could be widely used. As with all new methods to learn, the best way is to visit a lab who is well known with the method, and to learn it there to avoid all pitfalls. The short duration of the Uterotrophic bioassay makes a visit to another lab less complicated.

Member 4

The level of training required to conduct the Uterotrophic Bioassay and evaluate the gravimetric responses is reasonable for its wide use. More training and expertise is needed for an evaluation of additional endpoints (biochemical markers, histological changes).

Member 5

Yes. To perform the uterotrophic bioassay at least the personal need to know rat basic anatomy, must have practice in oral gavage – this is the most critical point in unaccepted deaths – and must have experience in removing tissues from animals. The previous things are thought to technicians in a few days, so the necessary expertise and training is very reasonable. Animal care is the standard, nothing special, weight measurement again is the most basic necessary practice in either biological or chemical laboratories. Just a good team-work of three or four people is enough.

Member 6

In light of the rather flat dose response curve of strong agonists and the resulting uterine weight increase, the level of training is crucial to obtaining relevant and reproducible results. The present international situation for availability of the necessary know-how and training does not readily support adoption of this protocol in any laboratory.

Member 7

Given the difficulties mentioned with some laboratories following protocols or submitting accurate data, further training may be warranted. The inclusion of acceptable performance criteria for the assay in the protocols / test guideline would be a way to address this.

Member 8

[Secretariat: No reply received from panel member on this point]

Observer 1

[Secretariat: No reply received from panel observer on this point.]

Responses for 20th October, 2003, teleconference

d. Are the necessary equipment and supplies relatively easy to obtain?
Member 1 Yes, another of the attractions of this assay.
Member 2 Yes, surgical instruments should be available in reasonably equipped laboratories. One item that should be considered is calibration of scales. Scales should be calibrated at least once per year.
Member 3 Since I have no practical experience with the Uterotrophic bioassay in the lab it is not easy to comment on, however, from the protocols it looks out not to be difficult for a lab to obtain all the necessary equipment and supplies to perform the Uterotrophic bioassay.
Member 4 All the equipments and supplies required for the gravimetric Uterotrophic Bioassay are easy to obtain.
Member 5 Any laboratory willing to perform uterotrophic assay should have all necessary equipment even animal work can be done in a standard air-conditioned laboratory not in continuous use. I think, easy and not expensive – in absolute terms**.**
Member 6 [Secretariat: No reply received from panel member on this point.]
Member 7 It would appear so.
Member 8 [Secretariat: No reply received from panel member on this point]
Observer 1 [Secretariat: No reply received from panel observer on this point.]

Responses for 20th October, 2003, teleconference

e. Is the method cost-effective, relative to the cost of conducting other *in vivo* assays?
Member 1 This test balances the costs of animal procurement and housing with the real value of an *in vivo* response and quite short exposure periods. While proponents of in vitro screens would argue that a transfected gene and reporter system would be more cost effective, nothing takes the place of a real response in a living system, so as a second- or third-tier test, this one is valuable and quite cost-effective.
Member 2 Yes
Member 3 I consider the method as cost effective, compared to other *in vivo* assays of longer duration. The Uterotrophic bioassay is a short-time screening assay, measures few endpoints, and there is not necessary with high-cost biotechnology equipment. However, on the other hand the bioassay is limited in endpoints studied.
Member 4 The method is cost effective, viewed either by itself or relative to other *in vivo* assays.
Member 5 Yes. We have experience in short-term and long-term (2 yrs) carcinogenicity assay followed by complete histological examination of a few hundreds of mice and rats… comparing to that this is very easy and cheap.
Member 6 [Secretariat: No reply received from panel member on this point.]
Member 7 Unable to comment.
Member 8 [Secretariat: No reply received from panel member on this point]
Observer 1 [Secretariat: No reply received from panel observer on this point.]

Responses for 20th October, 2003, teleconference

f. Is the time needed to conduct the Uterotrophic Bioassay reasonable?
Member 1 Would that ALL *in vivo* tests were this short. Yes, it is delightfully short.
Member 2 Yes
Member 3 The Uterotrophic bioassay is a short time screening assay, and as stated in the protocols, the bioassay can be performed during one weak.
Member 4 The Uterotrophic Bioassay itself is completed within a week; overall time including preparations (such as ovariectomizing animals and awaiting decline of endogenous hormones) is still reasonable.
Member 5 Yes, using any of the four protocols, results can be get fast.
Member 6 [Secretariat: No reply received from panel member on this point.]
Member 7 In the context of a short-term in vivo assay, yes.
Member 8 [Secretariat: No reply received from panel member on this point]
Observer 1 [Secretariat: No reply received from panel observer on this point.]

i. 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?

Member 1

The Phase 2 report shows that animal mortality can significantly impair the ability of the test to identify estrogenic compounds. It seems to be able to tolerate the loss of one animal, but losing more than 1 animal appears to weaken the test. I agree with placing the number of animals such that the test is not compromised critically by the loss of 1 animal (i.e., n=5 would be too few); the test needs to be sufficiently robust to be able to lose an animal and not be critically compromised. Six seems to be the right number.

Member 2

I was surprised that a sample size of 6 yielded relatively small standard errors about the mean responses in the test laboratories. This would suggest that an n of 6 yields adequate statistical power to detect a uterotrophic effect. My estimate of sample size would have been higher.

Member 3

The statistically power is considered better when using higher number of animals in each dose group, however, from table 9 it seems that 6 animals per group appears to be sufficient for detecting a 25-35 % increase in uterine weight following exposure to the weak oestrogen agonists in phase 2 study.

Following exposure to other weak partial oestrogen agonists, where a weak uterine response in achieved in the lower portion of the dose-response curve, it is more likely to assume that a higher number than 6 animals is necessary to obtain valid results.

However, to minimise the use of animals is in line with the EU chemical policy, where the goal is to reduce the number of animals in toxicological testing.

Member 4

The group size of six animals per dose group is appropriate.

Member 5

Yes.

Member 6

[Secretariat: No reply received from panel member on this point.]

Member 7

In relation to a group size of 6, yes (statistical considerations well-documented). However, in the context of continuing with both the immature and OVX models, then no. Given the inclusion of proper range-finding studies in immature rats prior to undertaking a definitive study (which should prevent the deaths recorded in some instances), then there is no justification for using a surgical procedure (ovariectomy) given that the conclusions of phase 1 (and then phase 2) were that the overall performance of the two versions of the bioassay was indistinguishable - both using the potent reference oestrogen EE and the five weak oestrogen agonists. Practical reasons of timing in relation to the immature rats are not sufficient to condone the use of the OVX model when an obvious refinement, the intact sexually immature rat model, is available. In the EU, if Directive 86/609/EEC (animal protection legislation) is properly enforced, the OVX model would not be permitted. There are also issues in relation to the incomplete removal of ovarian tissue with the OVX model.

Member 8

[Secretariat: No reply received from panel member on this point]

Observer 1

[Secretariat: No reply received from panel observer on this point.]

3rd December, 2003 – Second Set of Peer Review Panel Responses

This section records all comments from the Panel Members and Observers for the Uterotrophic Validation Peer Review on the second round of charge questions. These were discussed at the teleconference on Wednesday, 3rd December, 2003.

Test Method Data Quality and Sufficiency
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.
Member 1 On the surface, yes, the data are of sufficient quality. I have not plowed through the GLP statements, but if the GLP's are working as advertised, then it's down to trust, which is at the root of all science. Past that, there's not much else we can do except trust that each lab is doing the work honestly and with integrity.
Member 2 Yes, there were some problems but they did not affect the overall evaluation of the compounds tested.
Member 3 The data generated in both phase I and II are of good quality, i.e. more than sufficient! This is noteworthy in light of concerns by some panellists how more or less strict adherence to protocols and procedural variations might affect the outcome.
Member 4 Overall, yes the data are of sufficient quality. However, there are some cases where the study design was compromised and instances of "missing data". The ability of some laboratories to follow precisely the protocols and submit accurate data is questionable. The laboratories were requested to comply with GLP guidelines, but not all did so.
Member 5 My general impression is that adherence to the protocol was sufficient to allow for well-funded conclusions about the results. The few data that were not entirely clear do not affect the overall usability of the complete dataset.
Member 6 The generated data are of sufficient quality. The few instances of abnormal data are clarified satisfyingly. However, Kanno et al (2003) stated that ….*full GLP compliance was not a requirement for a laboratory's participation in the validation program, and several Laboratories did not perform their studies under GLP*. A judgement if the study has been performed in compliance with GLP requires an in depth review of all documentations of the participating laboratories. Kanno et al (2003), Environ Health Perpect 111:1550-8
Member 7 Overall, the data generated are considered of sufficient quality, although some laboratories did not perform the assay in compliance with Good Laboratory Practices. When suspect data was found e.g. in the phase 2 dose response study for NP exposure with protocol B (lab 20) , it was found that the diet influenced the results. One limitation in the phase-2 dose response study is that with protocol D only 2 labs participated for the weak agonists tested, and no negative control, such as DBP was included in the phase-2 study. For some of the weak agonists tested the lowest dose that induced a statistically significant increase in uterine weigh varied between labs e.g. BPA 375 and 1000 mg/kg with protocol A and 100 and 600 mg/kg with protocol B, which emphasis the need for good skill and lab experience.
Member 8 I think carefully performed experiments with strict documentation may prove the data quality obtained. Any laboratory procedures should be performed with detailed documentation and with care and precision.
Observer 1 There are indications some laboratories had some problems adhering strictly to the protocol.

Responses for 3rd December, 2003, teleconference

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?
Member 1 The VMG and the participating laboratories did an exhaustive job in conceiving of and performing and analyzing experiments that were designed to evaluate the ability of this study design to be run in many different laboratories in multiple countries and continents. As near as I can tell, ALL the data are summarized and presented describing the variation of the assay when trying to find a potent estrogen, and when finding weak estrogens, and estrogen antagonists. A thorough analysis of a couple of the most likely confounders is also provided in one of the manscuripts. So, you have an evaluation of how well it finds a strong positive, and the weak positives. The only thing missing is a description of how the assay performs in response to other steroids, like androgens or glucocorticoids; i.e., what is the false positive rate? The Owens and Ashby paper in Crit. Rev. Tox. mentions literature reports of such activity, and while I'll concede that this might be best evaluated by just a few laboratories, at SOME point it will be useful to get a sense of the penetrance of false positives across many labs. A small –scale evaluation will be necessary; a large-scale one would be nice but not necessary.
Member 2 Yes.
Member 3 The data are provided in sufficient detail to allow for an independent evaluation of results and assay performance.
Member 4 Yes.
Member 5 For the chemicals studied, the data are provided in sufficient detail. It is particularly interesting to note that the effective doses in the uterotrophic assay mirror those found in reproductive toxicity studies with different designs. However, in spite of the vast amount of data generated, I have reservations about the generalization of the results. Five positive compounds and one negative compound were tested, which is a low number in a formal validation. One cannot conclude from testing one negative about the possibility in general of false positives, compounds that may induce uterotrophy through a mechanism different from estrogenicity.
Member 6 The study design did not take the assessment of the performance sufficiently into account. Data on antagonists, negative compounds, data over time for more than two time points per lab. (for several substances), on coded doses over time and on the relation of negatives control over time are lacking, as well as data for a sufficient number of substances to assess the predictive capacity of the assays. An explicitly defined data interpretation procedure/ prediction model is lacking. It should have been complete by covering the death of animals and also the way to consider adverse events, e.g. loss of body weight. A final trial plan, which includes e.g. the objectives of the validation study,, the study design or the data analysis, was not included in the submission package if ever produced. It would have been substantial for following the reasoning underlying the validation study and how the performance was planned to be proven.
Member 7 Yes, the data is considered provided in sufficient detail to evaluate the results and performance of the Uterotrophic bioassay. However, one limitation is that the various labs did not test the weak agonists at all dose levels in the phase-2 dose response study.
Member 8 Yes.
Observer 1 Generally not. It would have been far more straightforward to present actual uterus weights and uterus weight change rather than relative weight change ratios. Graphics showing dose response curves with error bars would have been better than tables. A significant omission is that ratios are reported without baseline data. The presentation and analysis of the between laboratory reproducibility and the consequence of high between laboratory variability on the predictive capacity of the test is insufficient (see attached supplemental analysis).

Responses for 3rd December, 2003, teleconference

Test Method Performance

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?

Member 1

The range of compounds chosen was appropriate to show that estrogenic compounds would increase weight, but it was inappropriate to show that every compound that increases weight IS an estrogen. (All elephants are grey, but not everything that is grey is an elephant). It was nice to have a compound that was metabolically activated to a more potent agonist (MX). The selection criteria as summarized on pg 20 of the Phase 2 report are reasonable and appropriate. It's valuable to have a body of pre-existing data showing estrogenic effects at multiple levels, from receptor binding through to multi-gen studies. The main effect of this is to add credibility and a large dose of context to the results of the uterotrophic results under discussion. An additional benefit of this particular group is that there is one plant-derived phytoestrogen, one long-term environmental contaminant (DDT), and several in significant current use (BpA, BP, and DBP).

We can conclude from this selection that it will find estrogens, but a uterotrophic response alone will not tell us that the chemical in question is an estrogen; we need receptor-binding data to distinguish between an androgen and an estrogen.

Member 2

Yes.

Member 3

The test substances cover a wide spectrum of characteristics in terms of potency, metabolism (bioactivation/inactivation), and lipophilicity; thus the selection was adequate for investigating the usefulness of the UT as a screening assay.

Member 4

At present, the collated database on chemicals tested in the validation study is extremely limited, i..e. essentially it is EE (potent agonist), ZM (antagonist), DBP (negative control) and the 6 weak oestrogen agonists tested in phase 2. This is a major weakness when trying to take a definitive decision about the validity / regulatory testing application of the bioassay in a much wider "general chemicals testing" context. There is limited coverage of chemicals of different classes, and extremely limited coverage of test substances with different activities (even in respect to clear non-oestrogenic and anti-oestrogenic chemicals). Given the potential for false positives, and the manner is which tests can be used / "abused" in a regulatory context, I would have more confidence in the validity / interpretation of the bioassay data if the database could be extended to include several more relevant substances for which developmental toxicity / multi-generation study data are already available for comparative purposes.

Member 5

For the estrogen agonists, the five compounds tested show good results. We do not have sufficient data to decide about possible false positives. For the estrogen antagonists also only one compound was tested. In view of the fact that antagonists are specifically tested for their interaction with an estrogen-driven uterotrophic response I have less reservations about false positives among antagonists, although formally, only one compound tested is clearly less than ideal for a validation study.

Member 6

The mere number of test substances was not selected adequately to address the performance of the assay. As only one antagonist was tested at one time point, only the feasibility of the assay to screen for antagonists was shown. Furthermore, the specificity was not adequately considered, so that no conclusion about false positive can be drawn (lacking of an adequate number of neg. substances).

Responses for 3rd December, 2003, teleconference

Test Method Performance
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?
Member 7 Yes, the characteristics of the test substances selected were adequate to demonstrate the performance of the Uterotrophic bioassay for its intended use both for weak agonists and antagonists. However, maybe more than one negative substance (DBP) should have been included in the evaluation of negative results. Since humans are exposed to more than one weak oestrogen every day at low doses, this issue could have been addressed, since the focus on exposure to mixtures is an increasing issue.
Member 8 Selection was correct for the chemicals especially because also chemicals were selects which showed to have positivity in the previously performed uterotrophic assay. I think that in vivo testing is necessary for those (new) chemicals also where no direct receptor binding was proven because of the possible "metabolisation". In the VMG-NA it was decided that for in vitro tests no metabolizing system will be used (because metabolic modification produces a new substance). In these cases Protocol A (per os) application is necessary.
Observer 1 No. The overall number of test substances is too few. The number of negative test substances tested is too few (testing only one negative substance is ridiculous). See references (1-4) for guidance on test substance selection for validation studies. [these references have been inserted immediately below by the Secretariat for reader convenience] The number of chemicals used to assess the test's capacity to predict oestrogen antagonists is too few.

1. Balls M et al. Practical aspects of the validation of toxicity test procedures. Alternatives to Laboratory Animals 1995;23:129-47.
2. Bruner LH et al. Validation of alternative methods for toxicity testing. In: Marzulli FN, Maibach HI, eds. Dermatotoxicology. Washington, DC: Taylor and Francis, 1996:579-605.
3. Bruner LH et al. An investigation of new toxicity test method performance in validation studies: 1. Toxicity test methids which have predictive capacity no greater than chance. Hum.Exp.Toxicol. 21[305], 312. 2002.
4. Bruner LH et al. An investigation of new toxicity test method performance in validation studies: 3. Sensitivity and specificity are not independent of prevalence or distiribution of toxicity. Hum.Exp.Toxicol. 21, 325-334. 2002.

Responses for 3rd December, 2003, teleconference

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 substances?
Member 1 A priori, it would seem that ANY chemical is a potential candidate for being tested in the Uterotrophic Bioassay (UB). However, on further reflection, not every compound will likely be put through this assay: The Kanno phase 1 paper in EHP describes on the first page that the UB is meant to be part of a tier of increasingly complex and definitive assays, with the first step in this process being some sort of SAR method, followed by a receptor binding assay, and (in a logical world) only those compounds that give some signal in both of those will likely go on to further testing in the UB. This will likely limit the list of potential candidates to those with some aromatic function, and within a certain molecular weight range. While it might have been ideal to have included some weak estrogens that are structurally very different from the majority of those tested here, or perhaps some strong and weak **androgens** this assay is allowed a little lattitude by virtue of it being used in conjunction with SAR, ER binding, and multigen data, so any erroneous signals that come through will be checked and checked again.
Member 2 Yes, but it is hard to say that all estrogenic substances in the future will be detected by the uterotrophic assay. Results from other assays, such as the MCF-7 cell proliferation assay, should continue to be considered when those data are available. For example a recent study by Howdeshell, et al. (EHP, 111(9): 1180-1187, 2003) suggests that Bisphenol A can be detected at lower concentrations using the cell proliferation assay than using the uterotrophic assay.
Member 3 The compounds chosen represent typical environmental chemicals for which the test is proposed to be used. Interestingly, the test method also performs well with "surprises", such as cadmium (Johnson MD et al, 2003, *Nat. Med.* 9 (8), 1081-4 and comment by Safe S (2003) ibid, 1000-1001).
Member 4 I was unable to find any definitive statement in the test protocols indicating which types of substances the bioassay is meant to be used with. If the assay is meant to be generally applicable to testing any chemical to determine whether it demonstrates oestrogenic (or anti-oestrogenic) activity, then I consider that the underpinning database is still insufficient (see comment above), and the proposed use of the assay in a battery or testing strategy approach is still insufficiently well-defined, to enable a decision about its overall validity (predictive performance) for regulatory testing purposes to be made at this stage. Insufficient substances have been tested in the actual validation study to draw any conclusions about limitations on the applicability of the bioassay to testing certain types of chemicals. However, overall experience with the bioassay outside of the validation study, suggests general applicability for oestrogen agonist and antagonist activities.
Member 5 See above. All depends on how this assay is used in a tiered approach. For compounds that have been tested in an in vitro estrogen receptor activation test, the data strongly suggest that the uterotrophic assay will detect a uterotrophic effect. For compounds for which estrogenicity in vitro is unclear, it cannot be excluded that the uterotrophic assay gives a false positive response.

Responses for 3rd December, 2003, teleconference

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 substances?

Member 6

The selection does not represent the types of substances for which the test method is proposed to be used.

The test chemicals have been selected i.e. due to their positive in vitro results (Kanno et al, 2003; Kanno et al., 2003a). However, the advantage of the Uterotrophic test in comparison to in vitro methods is the presence of metabolism and other parameters related to toxicokinetic. Since no chemicals have been selected that needed a first pass effect to be activated the test cannot claim to be able to detect also "proestrogenic" compounds. The same scenario holds true for chemicals that are very quickly eliminated before they start to act.

The negative compound has been selected due the neg. activity in the uterine estrogen receptor binding study and in vitro toxicological studies, including gene activation profiles (Kanno et al, 2003). However there are some evidences that also DBP has a weak oestrogenic potential in vitro (Harris et al., 1997; Jobling et al., 1995) and in vivo (Ohtani et al., 2000; Tollefsen et al. 2002). This should be at least discussed.

The experimental set up does not allow to provide more detailed information of a chemical with specific modes of action such as tamoxifen. Due to the lack of an endogenous oestrogen source the partial agonist and antagonist potential of a chemical cannot be identified.

Kanno et al. (2003) Environ Health Perpect 111:1550-8
Kanno et al. (2003a) Environ Health Perpect 111:1530-49
Harris et al. (1997) Environ Health Perpect 105:802-11
Jobling et al. (1995) Environ Health Perpect 103:582-7
Ohtani et al. (2000) Environ Health Perpect 108:1189-93
Tollefsen et al. (2002) Mar Environ Res, 54:697-701

Member 7

Since the Uterotrophic bioassay is a robust screening method, there is not considered to be great limitations on the applicability of the bioassay to certain test substances. However, concerns arise when test substance only activate the expression of early oestrogen regulated genes not involved in cell division.

The bioassay has also limitations in detecting weak oestrogen antagonists.

Member 8

My opinion is that there is no "general method" in biology, in endocrinology, etc. The in vivo uterotrophic assay – using both po and sc application – is able to detect even weak estrogen agonists and antagonists if tested in pure form. In multicomponent systems I would require the testing of the components one by one and the complete mixture together. For limitations see above this page.

Observer 1

No. The reference set of test substances contained an insufficient number of negative test chemicals (1-4). It is premature to make general recommendations regarding the utility of the test for use in as a screen for uterotrophic effects of chemicals. [references are below previous question – Secretariat]

Responses for 3rd December, 2003, teleconference

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 estrogen agonist and antagonist activity? If not, why not?
Member 1 The biggest question, after "Will this test find estradiol as positive?" is "How sensitive and consistent is the test at finding weak estrogens?", because most of the new compounds that will be found by this test are very likely to be weak. I thought this validation effort was outstanding at really exploring the variation across labs using a number of weak estrogens from a modest variety of classes. There was significant exploration of the LOAEL, and thoughtful discussion of the meaning of the variation across labs for this measure. While there is more variation than one would ideally hope for, there also is not more than one would expect to see for something like this. That is, ANY multi-lab and multi-national effort will have significant variability, and this is no exception. The good news is that the majority of the assays were consistent with each other in reporting relative potency of the compounds for increasing uterine weight (again the androgen issue means that we're finding uterotrophic compounds, we're not finding only estrogens).
Member 2 Yes.
Member 3 The dose response experiments indicated that the UT-assay can distinguish between strong, intermediate and weak agonists and/or antagonists and therefore adequately demonstrate the toxicological performance. From a toxicologists point of view it seems highly desirable to obtain more data points, including some in the lower range of a dose response curve to allow application of benchmark modelling in the future (see comment below on allocation of animals to dose groups).
Member 4 *The data on the weak agonists summarised in Table 12 of Owens & Ashby (2002) show the predictivity of the uterotrophic bioassay for these 6 substances. I would like to see this table extended to include other data that may already be available for additional chemicals tested outside of the validation study, to support a weight of evidence.* It is clear that the uterotrophic bioassay is not necessarily predictive for the dose at which oestrogen-related responses were observed in developmental and multi-generation studies. I do not consider this unsurprising, given the likely different sensitivities of the assays. This does not negate its use as a predictive tool if this can be adequately established.
Member 5 Dose-responses are sufficient for supporting the conclusions about the (anti)estrogenicity of the compounds tested.
Member 6 Assuming the toxicological relevance of the endpoint, there is only one antagonist, which was tested at one time point and only in two concentrations. Therefore, the toxicological performance for antagonists remains largely unknown. Furthermore, its use in screening cannot be assessed, as the assays sensitivity has not been addressed in sufficient detail.
Member 7 Overall the use of the test substances in dose response experiments were adequate to demonstrate the toxicological performance of the Uterotrophic bioassay, however, one limitation is that the various labs did not test the weak agonists at the same dose levels that were recommended.
Member 8 Yes, it was adequate.
Observer 1 No for two reasons: 1) The dosing regime was specified to the laboratories. This procedure controlled an important source of between laboratory variability in a study designed to test the between laboratory reproducibility of results and predictions! 2) There were no negative chemicals tested in the dose response studies.

Responses for 3rd December, 2003, teleconference

Comment on the adequacy of the statistical/analytical methods used to evaluate the performance of the Uterotrophic Bioassay.
Member 1 Joe Haseman is one of the world's best statisticians, and his grasp of how to handle numbers, and what tests to use for which purpose, is unparalleled. This is really out of my area, I can only say that it appears that the appropriate things were controlled for, and the right things were analyzed for (variance, outliers, confounders). While I would have liked to have seen some indication of significance for measures other than just the organ-to-body weight ratio, it appears that the statistics were done appropriately.
Member 2 Seems fine to me.
Member 3 All statistical methods used were adequate to analyse positive/negative responses at a given dose level and to evaluate the overall performance.
Member 4 *The data analyses were undertaken independently of the participating laboratories by professional statisticians experienced in analysing data sets submitted by very many laboratories. A more standard data submission process, agreed between the statisticians and the participating laboratories in advance of the testing part of phase 2 of the validation study (and subsequently adhered to), might had facilitated the overall data management.* Statistical methods and considerations have been well-documented. They should also be incorporated into the revised test protocol(s) for the bioassay, and in any future test guideline.
Member 5 No specific remarks, the analysis appears adequate.
Member 6 The statistical methods are adequate. As the ANCOVA includes the bodyweight, which can be effected by toxic effects of substances, the impact of this dependency in the model could have been discussed. The primary results of the ANCOVA could have been reported. A clear statement of the PM/DIP, which should be complete with regard to adverse events, would have been helpful. Due to the overall study design, i.e a limited set of substances, a final assessment of the performance in terms of predictivity cannot be done. Although not asked for but considered a very important: It remains unclear how the study was designed and who was involved. A document explaining how the aims of validation are addressed in the study design should have been produced/provided.
Member 7 No comment.
Member 8 I am not experienced in statistics over SD, SEM, or regression analysis. As I see careful analysis was performed in many points of view.
Observer 1 I have received data tables, but have not yet received the statistician's reports. It would be helpful to see the statistician's assessment of results from the study. The data analysis provides an inadequate assessment of the between laboratory variability in weight change and the consequences of the observed variability on the predictive capacity of the test. Between laboratory CV of uterus weight change runs in the tens to hundreds of percent. The effect of this high level of between laboratory variability on the predictive capacity of the test has not been adequately addressed (see attached supplemental analysis for an example of what is needed). The use of the ANCOVA procedure is most likely inappropriate since a spot check of the data suggests the assumptions (regression of uterus weight on body weight and parallelism of the regression lines between doses (see Sokal and Rolf ((5)) are not met. (see figure 8 in attached supplemental analysis. It should not be assumed that labs who did not test all doses would have achieved statistical significance. Their reported predictions should be classified negative.

Responses for 3rd December, 2003, teleconference

Based on the Submission Package, are results of Uterotrophic Bioassay relevant and predictive for possible oestrogen agonists and antagonists?
Member 1 The data which are reviewed in the Background Document (pp102-107) do a good job of reviewing the estrogenic responses of these compounds. It is clear from this review that these compounds clearly DO act like weak estrogens in vivo, based on a number of different endpoints (the most sensitive of which generally is developmental). From this comparison, one might conclude that developmental exposure would be a preferred design, but comparison shows that uterotropism occurs within an order of magnitude of the developmentally-active doses, which suggests that the UB will effectively identify uterotrophic compounds for further testing, and prove activity in vivo, which is the intent of the test. It has not been shown to distinguish estrogens from androgens or glucocorticoids.
Member 2 Yes.
Member 3 The results obtained in phase I and II show already that the bioassay is relevant and predictive for oestrogen agonists and antagonists. In addition to many papers cited in the background document, also the long and successful use of this bioassay in the screening of pharmaceuticals with such hormonal activities can be considered in this context.
Member 4 The information in the Submission Package generally provides support for the hypothesis that the results of the uterotrophic bioassay are predictive for potential strong and weak oestrogen agonists. The performance/relevance for anti-ostrogens cannot be properly assessed at this stage on the basis of the data for a single antagonist in phase 1. The database needs to be expanded (see earlier comments), to include a couple of oestrogen antagonists and a couple of negative substances as a minimum, to fully address the predictivity of the assay for possible oestrogen agonists and antagonists. These could be data from a single laboratory (reproducibility has already been fully assessed) obtained using the current bioassay protocol(s).
Member 5 Yes, for the specific compounds tested there is a good correlation with other in vivo data. But see remarks above about false positives and relatively few chemicals tested.
Member 6 The Uterotrophic Bioassay seems to be basically predictive for oestrogenic agonists, although the number of substances is limited. Again, information on its specificity as well as for antagonists is not sufficient to assess the relevance and predictivity in this sense. It needs further discussion if the test is meaningful for weak oestrogen agonists since the used dosage were in the range of toxic concentrations. The deaths in different experiments as well as other signs of toxicity have not been taken into account. The interpretation of the results will be even more difficult if one has to take into account the presence, in an intact female organism, of oestradiol (i.e. compounds with partial agonists activity like tamoxifen).
Member 7 Overall, the results of the Uterotrophic bioassay were relevant and predictive for the weak agonists tested when the uterotrophic data was compared with available reproductive and developmental toxicity data. For BPA an increase in uterine weight was observed in the phase-2 dose response study, however, there was some variability between the labs concerning the actual dose at which statistically significant was first achieved. Other endpoints for measuring an oestrogen mode of action such as mammary gland dysegenesis and luminal epithelial cell height are reported to be more sensitive for BPA. Effects on these endpoints are indicated to be induced at BPA doses lower than the BPA doses inducing oestrogen mediated effects in multiple generation studies. Since only one oestrogen antagonist was tested, limited information is available concerning the relevance of testing oestrogen antagonists especial weak antagonists with the Uterotrophic bioassay.

Responses for 3rd December, 2003, teleconference

Based on the Submission Package, are results of Uterotrophic Bioassay relevant and predictive for possible oestrogen agonists and antagonists?
Member 8 Yes, predictive and relevant if using both po and sc methods. Estrogen antagonists with weak activity is hard to detect but I do not know if weak antagonists (with approximately two orders lower activity) had any biological significance for humans. Their persistent appearance in the Nature (water, soil) may effect the development of the non-mammalian species. This is another section of EDTA.
Observer 1 Relevance: I defer to the endocrine experts on the committee to make the relevance assessment. Predictive capacity: the consequences of the high between laboratory CV of uterus weight change and in controls needs further assessment. See attached supplemental analysis for example of the type of analysis that should be done for the full study.

Responses for 3rd December, 2003, teleconference

Does the Submission Package adequately support the utility of the method for regulatory use in hazard assessment of chemical substances that may have the potential to act as oestrogen agonists and/or antagonists? If not, why not.
Member 1 The Submission Package shows that a large group of laboratories will obtain generally the same answer when they test the same compound using the same method. These data do support the use of this test to identify compounds that may **increase uterine weight** (which should be distinguished from estrogenic, because androgens will increase uterine weight), and when used **in conjunction with receptor binding tests**, can identify weak and strong estrogens. This test alone should not be used to assess risks from human exposure to these compounds.
Member 2 Yes, but this should not be interpreted that only the uterotrophic assay is effective nor should it prevent continued research on assays.
Member 3 Data given in the Submission Package indicate that the UT-assay is useful for detecting chemicals that are potential oestrogen agonists/antagonists. This is further supported by additional published data on other (environmental) chemicals with such hormonal activities. The method is thus adequate for regulatory use in hazard assessment. Again: From a *toxicologists* view it is highly desirable to obtain more data on dose responses, namely in the lower range of a curve to allow application of benchmark modelling in the future (see comment below on allocation of animals to dose groups). This view may not be shared by others, as also indicated by ongoing discussions on the regulatory use of data from this bioassay (e.g. merely qualitative or quantitative as well).
Member 4 *The underpinning database needs to be supplemented (see previous comments).* Given the probability of false positives (issues linked to specificity), there is a clear need to establish and agree defined processes for data acceptance and interpretation, and the way in which the assay would be used in any test battery/strategy approach (weight of evidence assessment) before it is implemented for routine regulatory use for identifying potential oestrogenic-related hazard.
Member 5 One has to consider the additional contribution of this test over in vitro estrogen response tests, and weigh this against possible disadvantages. The main additional value of the uterotrophic assay is its incorporation of aspects of ADME, its main disadvantage is the uncertainty about false positive outcomes. With ADME being critical, it should be stressed that only the relevant route of human exposure should be tested, which in most cases would not be subcutaneous administration (although this being a hazard characterization one could argue that the most sensitive route would be indicated, but this would take away the advantage of the ADME argument). In a tiered approach, the Background Review places the test between in vitro tests and more elaborate in vivo tests, for prioritization of further testing. I agree that this is the optimal place for the uterotrophic assay. One could however question whether the test is necessary if one has in vitro estrogenicity data and information about ADME from other studies, which may often be the case. In such cases, in view of reduction of animal use it would not be appropriate to do a uterotrophic assay.
Member 6 The study can only claim to identify direct acting oestrogenic agonists. Due to the lack of a sufficient number of negative reference test chemicals the predictive power of this test cannot be judged. Nothing can be said about the uterus growth that is stimulated by a different mode of action such as androgens, progestin and growth factors. A positive result in the uterotrophic assay cannot lead to the conclusion that the tested chemical is an oestrogen agonist. This can only be done in combination with in vitro test but since the advantages of the uterotrophic tests (metabolism and toxicokinetics) have not been used in the validation study there is no need to use the uterotrophic test for regulatory use at this stage.

Responses for 3rd December, 2003, teleconference

Does the Submission Package adequately support the utility of the method for regulatory use in hazard assessment of chemical substances that may have the potential to act as oestrogen agonists and/or antagonists? If not, why not.
Member 7 An increase or decrease in the uterine weight implies that there has been an exogenous source of oestrogen/antioestrogen exposure. Such data can be used in hazard assessment, however, for deriving N(L)OAEL values to be used in the risk characterisation for workers, consumers or indirect exposure from the environment the NOAEL values have to be derived from more definitive studies i.e. multi-generation studies or developmental studies.
Member 8 Yes, in case of testing estrogen antagonists I would recommend more concentrations of the estrogen stimulation. I mean for oral administration one dose over 3 microgram/kg/day and for subcutaneous administration lower than 0.3. microgram/kg/day in case of testing low activity (weak) estrogen antagonists.
Observer 1 No. Additional analysis of the results should be conducted. The poor between laboratory reproducibility may limit the utility of the test. This finding needs further assessment. The appropriateness of using ANCOVA should be checked.

Responses for 3rd December, 2003, teleconference

Determination of Test Method Reliability (Repeatability/Reproducibility)

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

a. Taking into account the need to balance the use of resources and the generation of data, comment on the adequ inter-laboratory reproducibility of the test method?

Member 1

Using 21 laboratories to test inter-lab variability constitutes, in my opinion, a sufficient test of the robustness of the assay less variance, with little clear pattern. In short, the variance between labs is about what we would expect. In the context of using this variability should not be limiting. As long as no regulation (i.e., acceptable environmental exposure levels) will result f variability should not cripple the use of the assay.

Member 2

Yes.

Member 3

Yes. (and below - comment on allocation of animals to dose groups).

Member 4

Reproducibility has been "over-assessed" (considerably more laboratories than actually needed to make an objective ass reproducibility), thereby complicating the study unnecessarily and not making most effective use of the resources available. The reproducibility / transferability at the expense of sufficient chemicals being tested in the validation study to fully assess the prec bioassay.

Member 5

Yes, the evidence on reproducibility is quite convincing from the data.

Member 6

Intra- and inter-laboratory reproducibility information is unbalanced: On the one hand information on intra-lab reproduci concentrations) identical experiments were reproduced in time. Only for EE (2 concentrations/protocol) more than two ru information on antagonists and non-oestrogenic substances. Lab related effects, e.g. due to their experience, are not discussed protocols and therefore also the reliabilities of the results differ. Three time points for several substances and/or concentrations w

On the other hand inter-laboratory reproducibility was considered excessively, but unbalanced with regard to protocols. An saved a lot of resources (i.e. 7200 animals in the whole validation study) by balanced approaches and simplified comparison. reproducibility, optimally with a confidence interval, would have allowed for a more comprehensive and comparable evaluation.

Member 7

In the phase 1 study with the reference oestrogen EE there was acceptable agreement among laboratories (n = 19) with re significant responses were obtained for a given protocol.

In the phase 2 dose response study the reproducibility of the dose response among laboratories (n = 20) was good, despite e.g., strain, diet, housing, protocol, bedding, vehicle used, and technical experience. However, there was some variability statistically significance was first achieved, especially for BPA with protocol A and B.

As mentioned earlier, the different weak oestrogen agonists was only tested in 2 labs with protocol D which limits the inter-l

Member 8

As I wrote I am not a statistician, but reading the phase 2 report, I feel that adequate and careful analysis was performed.

Observer 1

The consequence of the between laboratory variability observed in the study on the test's predictive capacity has not yet bee above and attached supplemental analysis. It should not be assumed that laboratories who did not test all doses would have reach

Responses for 3rd December, 2003, teleconference

b. Taking into account the objective of providing a Test Guideline that can be used widely and internationally 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)?

Member 1

The paper evaluating the level of phytoestrogens in the diet was very useful, and even more-so as it sheds some possible explanatory light on previous reports unrelated to this current effort. The authors of these papers chose not to evaluate the variability induced by bedding or diet because of the essential uniformity of the response. Given more time and resources for such an exam, it might have been instructive to have examined the responses from each lab, to determine if one type of bedding consistently was associated with increased or decreased responses, for example, even within the "normal" range. The lack of such an evaluation is not a problem, but some interesting and possibly useful relationships might have been uncovered with such an evaluation.

So, there ARE data to show that diet composition may affect the responses; there are no data to support the importance of bedding or vehicle, but these were not strictly evaluated.

Member 2

Yes, there was remarkable consistency.

Member 3

Many panel members (including myself) consider the flexibility of the Test Guideline allowing e.g. either use of immature or ovariectomized rats as an advantage. But, in light of data presented recently by Ashby (at the Workshop on Low Dose Effects of Endocrine Active Compounds. 20.-22. November 2003, Berlin) that high caloric intake can apparently advance puberty, the immature rat version seems less robust to interference unless food intake is restricted and/or age at treatment is kept narrow (early enough before puberty).

Other conditions:

Doses: This panelist, if given a choice how to allocate a given number of animals (e.g. n = 30) to treatment groups, would prefer to have 5 dose groups (of n=6) rather than 3 dose groups (of n=10) simply because a design with more than just 3 doses would make the results (of this and any other assay) more relevant for human hazard assessment.

Route: Subcutaneous and oral dosing are both suitable for hazard, although oral administration is more relevant in quantitative terms.

Member 4

The potential major sources of variability have been identified and investigated in parallel with the validation study and the outcomes reported. These findings need to be documented in the standardised protocol(s) and referred to in any test guideline, such that they can be controlled and any potential impact on data quality and interpretation minimised, by the testing laboratory.

The inclusion of a couple of standard / reference chemicals in any future test guideline, accompanied by criteria for data acceptability, would provide a means to determine that a laboratory was performing the assay correctly and that any flexibility due to protocol differences was not significantly impacting on data interpretation and the conclusions made.

Member 5

Except for the high phytoestrogen diets, which were clearly addressed in the study, there do not seem to be major factor affecting the results, given that the protocol is used properly.

Member 6

There is no evidence that these variables affect the reproducibility. However, for an optimal assessment of these parameters the study should have been designed taken these variables into account. E.g. 10 of 14 labs have taken corn oil as vehicle and only 1 lab has used sesame oil 1 lab has used peanut oil and 2 labs have used olive oil. The variables should be more balanced in order to do an optimal evaluation. Nevertheless, an analysis showing the effect of these variables on the parameters variability and thus on the sensitivity is lacking. Additionally, an analysis linking the level of experience of labs, which might be confounding with regard to sensitivity and comparison of protocols, could have been useful.

Responses for 3rd December, 2003, teleconference

b. Taking into account the objective of providing a Test Guideline that can be used widely and internationally 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)?
Member 7 The levels of phytoestrogens in the diet may influence the reproducibility and sensitivity of the test, especially for immature rats which eat more on a body weight basis than ovariectomised rats. The level of phytoestrogens in the diet should be controlled. The technical skill may also affect the reproducibility and sensitivity of the test.
Member 8 No real strain-, bedding-, vehicle-specific differences were found, in case of testing really weak estrogen agonist the phytoestrogen content should be kept low.
Observer 1 First question is unclear. Second question: Yes, the causes of high between laboratory CV of uterus weight changes (tens to hundreds of percent) need to be assessed.

Responses for 3rd December, 2003, teleconference

c. Was the reproducibility of the test method adequately evaluated using coded (blinded samples)?
Member 1 The VMG did a good job in setting up the evaluation by using blinded samples and providing a thorough and specific protocol. This is the best way to perform a rigorous evaluation of a method like this. It addresses the question: what is the inherent variability of a method across labs when everyone is being told to do the same thing? This is the base from which all evaluations must begin.
Member 2 Yes.
Member 3 Yes.
Member 4 Yes. Test chemicals were selected, coded and distributed independently of the participating laboratories.
Member 5 Yes.
Member 6 At least information on two tests on blinded samples should be carried out to assess the reproducibility to a sufficient extend. Thus the intra-laboratory reproducibility is not satisfyingly shown. Additionally, the inter-laboratory reproducibility could be biased by the laboratories experiences and/or for the protocols.
Member 7 The phase 2 coded single-dose studies were performed in 16 labs, and overall for each protocol, the mean relative increase in uterine weight was reproducible within and among labs for both the dose-response and coded single-dose studies with each test substance. However, one limitation in the evaluation of the results from protocol A was the high mortality observed with this protocol, that decreased the power of the test. Furthermore, a decrease in body weight gain was observed with various protocols and agonists, indicating that the maximum tolerated dose had been exceeded.
Member 8 Yes.
Observer 1 The data suggest the between laboratory reproducibility is poor (between laboratory CV tens to hundreds of percent) see previous comments. The effect of poor between laboratory reproducibility needs to be fully assessed.

Responses for 3rd December, 2003, teleconference

d. Considering the variability inherent in all chemical and biological test methods, are the results obtained with the Uterotrophic Bioassay sufficiently repeatable and reproducible?
Member 1 In short: Yes. The longer I am in science, the more impressed at the lack of replicability there is in any data set… i.e., how "squirrely" the changes are that we report. The data set reported by this group is no surprise. There is greater variation than we would all hope for, but not really an exceptional amount. This variation is neither better nor worse than other large data sets that are widely available.
Member 2 Yes.
Member 3 Yes.
Member 4 *Yes. Overall, reproducibility appears to be satisfactory.* "Normal" assay variability may underpin the generation of "false positives" and it is this issue that has to be addressed through provision of guidance on data acceptance/interpretation in relation to an overall assessment of the oestrogenic effects of a test substance.
Member 5 Yes.
Member 6 The basic biological variation in terms of CV is low for a biological test method. If the reproducibility holds true over time and over substances, it seems to be sufficient.
Member 7 Overall the Uterotrophic bioassay is considered a sufficiently repeatable and reproducible due to the test method's robustness.
Member 8 Yes.
Observer 1 Perhaps not. As noted above the consequence of the between laboratory CV in uterus weight change on the predictive capacity of the test needs further assessment.

Other Considerations

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:

g. Are there complications or limitations that have not been addressed by the protocols?

Member 1

As long as the necessary precautions are made for food and water, it should be transferable. Water, in particular, is not addressed in this protocol. (Now, one could argue that most estromimetics are lipophilic, and therefore unlikely to be in water at biological meaningful levels). It We should simply recognize that this is an area that has not been strictly evaluated, and is a possible vulnerability.

It would be useful to specify an acceptable range of body weights and control uterine weights for specific age and weight of rats, so that a lab which was implementing this method for the first time would have an external reference that it could use to assure that its baseline biology was in line with the rest of the world. An uterus already quite stimulated by phytoestrogens has less "room" to respond to weak estromimetics, thereby leading to false negative conclusions… helping to avoid that is the job of the people setting out the final protocol.

Member 2

Seems to be fine to be used in a variety of laboratories.

Member 3

(As mentioned above:) Data presented recently by Ashby (at the Workshop on Low Dose Effects of Endocrine Active Compounds. 20.-22. November 2003, Berlin) indicate high caloric intake can apparently advance puberty in the immature rat. Thus, recommendations to restrict food intake and/or on timing of treatment should be considered.

Member 4

Panel member has addressed this and the following question together.

The example protocols provided have not been optimised/standardised and do not contain all necessary details (see comments from various PRP members during the first stage of the review process). However, experience suggests that the assay is transferable between competent laboratories; several of the issues that arose during the validation study appear to be individual laboratory-specific and due to lack of adherence to the protocol and/or other guidance provided.

Member 5

I cannot identify other complications.

Member 6

The protocol does not state how to analyse the data. Therefore, also no guidance how to take the death of animals into account in the analysis is given. The body weight is not adequately taken into consideration since chemical has often been applied in toxic concentration. Additionally, the protocols are not suitable to be applied to real-life testing situations, this problem has not been addressed at all.

The dosage selection for unknown chemicals has not been described.

The relation between food consumption and body weight has not been taken into account. E.g. corn oil gavage provides additional calories suppressing subsequent food consumption, an observation reported previously (Ramirez and Friedman, 1983)

If measured daily food consumption" means "mean food consumptions" since cages contain more than one animal. Since oral gavage as route of administration has been used historically does not mean that it should be continue to be used. This route of administration is very stressful for the animals and it can influence the results (see point 17 pag.105 phase I: It has been stated that stress-increase hormonal control). Several reports have documented cases in both animals and humans, of oesophageal, gastric perforation and dyspnea, during oral gavage (the result of about several deaths in protocol A in this validation study can be explained also by one of this unforeseen events). Taking together all these observations suggest that at least an improved oral dosing technique should have been taken into account.

Responses for 3rd December, 2003, teleconference

Other Considerations

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:

g. Are there complications or limitations that have not been addressed by the protocols?

Member 7

It should be included in the protocols that in the immature assay pups outside a 35 to 50 g body weight should not be used on pnd 19 since pups smaller than 35 g may not respond as well as pups larger than 35 g whereas pups larger than 50 g may begin to secret E_2 by day three of the assay when they reach or exceed a body weight of 60 g.

Due to the appearance of false positive and false negative with the Uterotrophic bioassay, the need for clear criteria for data acceptance is necessary e.g. maxima for acceptable vehicle control uterine weights, and how to interpret a modest increase in uterine weight (20 to 40 %).

It should also be included in the protocol that when unknown substances are tested a reference substance e.g., EE and a known weak agonist should be included in the test.

Care should also be taken not to exceed the maximum tolerated dose, to reduce animal pain and mortality.

It also should have been included in the protocols that the level of phytoestrogens in the diet should have been controlled.

Member 8

I did not find any.

Observer 1

As noted earlier, the protocols evaluated in this program are developmental protocols not yet developed sufficiently for assessment in a state-of-the-art validation study. Notes found in the submission package and summary data files note problems laboratories had in following the protocols.

Responses for 3rd December, 2003, teleconference

h. Considering the need to employ the Uterotrophic Bioassay internationally, can the test method be readily transferred among properly equipped and staffed laboratories?
Member 1 It appears that it can, given the correct identification rate of 94% of the participating labs.
Member 2 Yes.
Member 3 The experience with the outcome of both phase I and phase II studies indicates that the test method can be readily transferred internationally among properly equipped and staffed laboratories (regardless of whether these use GLP or not).
Member 4 **Panel member has addressed this and the previous question together.** The example protocols provided have not been optimised/standardised and do not contain all necessary details (see comments from various PRP members during the first stage of the review process). However, experience suggests that the assay is transferable between competent laboratories; several of the issues that arose during the validation study appear to be individual laboratory-specific and due to lack of adherence to the protocol and/or other guidance provided.
Member 5 No problem.
Member 6 If the SOP s have been improved: yes.
Member 7 Yes, the method is robust and reproducible.
Member 8 Yes.
Observer 1 The high between laboratory CV in uterus weight change suggests there are transferability issues. The cause of the high CV to be understood before this question can be answered.

Responses for 3rd December, 2003, teleconference

i. Is there any other information that should have been added to the Submission Package, published or un-published?
Member 1 I can think of no omitted information. Additional studies (or literature-based conclusions) will be required to separate estrogenic activity from the uterotrophic activity of other steroids.
Member 2 No.
Member 3 No comment – need more time to contemplate this point.
Member 4 *Is it possible to expand Table 12 (Owens & Ashby, 2002) / Table 36, p. 111 in the DBR document, to supplement it with data for other relevant substances tested outside of the validation study but with comparable test protocols? Data from a single laboratory would be fine. The data would need to be made available for peer review. This would greatly facilitate a judgement on the overall validity (predictive performance) of the rat uterotrophic bioassay for correctly identifying oestrogen agonists and antagonists.* All supporting data/information requested by the PRP members that was not initially provided in the Submission Package has been distributed in a timely manner by the Peer Review Coordinator (e.g. statisticians reports, raw data, etc.).
Member 5 No.
Member 6 The contrary results of the negative compound DBP was missed in the submission package. This point was addressed in the 2nd question of test method performance.
Member 7 Not, as far as I can see.
Member 8 No, perhaps a few more words about the function(s) of ER-alpha.
Observer 1 Yes. See previous comments.

9th February, 2004 – Summary Peer Review Panel Responses and Overall Discussion

This section records all comments from the Panel Members and Observers for the Uterotrophic Validation Peer Review on the summary charge questions. These were discussed at the teleconference on Monday, 9th February, 2004.

Has 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?*

Member 1

Most aspects of the Uterotrophic Bioassay have been thoroughly evaluated: we know that when specified doses are tested, and analysis is performed by a central laboratory, most of the agonists will come up positive. Only a single proof-of-concept study was performed with an antagonist, and while the concept was proven, the actual performance of the test in identifying antagonists was not demonstrated. The bioassay protocol still needs to specify methods of setting doses, statistical methods, methods for identifying estrogen antagonists (relative doses of estrogen and the antagonist, and timing of administration), and criteria for identifying that a compound is or is not estrogenic. When these are addressed, and only when used as part of a multi-step process that involves receptor-binding studies and fertility assays, I believe this test would be ready to be used to help identify estrogenic compounds.

Member 2

For estrogen agonists, yes, on the basis of a limited set of agonists tested plus existing knowledge about the mechanism of uterotrophy. For antagonists, too few compounds have been tested to make a final statement. Remaining issues are however specificity (e.g. the uterotrophic response to androgens) and sensitivity (e.g. guidance on dose-response testing, statistical methods and use of positive and negative controls).

The answer to this question should also be related to the place of the test within a tiered approach. For example, the uterotrophic can be a useful in vivo follow-up of in vitro estrogen receptor binding and activation assays. On its own, the uncertainty about specificity and sensitivity limits the predictivity of the assay.

Member 3

[no answer here – Secretariat]

Member 4

Overall, the Uterotrophic bioassay has been sufficiently evaluated by the OECD validation program to support its proposed use for screening the potential of substances to act as oestrogen agonists or antagonists *in vivo*. However, the use of only one negative substance, DBP, may limit the validation of the Uterotrophic bioassay. At least 2 to 3 negative substances should have been included to better quantify false positive.

Sine the Uterotrophic bioassay is an *in vivo* screening assay for oestrogenic agonists/antagonists, a positive response is not of itself an adverse effect but could be indicative of other or adverse properties of the chemical. A positive response in the Uterotrophic bioassay suggests the need for the substance to advance to reproductive and developmental testing for adverse effects.

Member 5

Yes, over-all the Uterotrophic Assay has been adequately characterized. This is especially true for estrogen agonists but is weaker on the antagonists.

Member 6

The Uterotrophic Bioassay studies conducted within the OECD Program (as well as additional published data, reviewed in the background documentation) clearly justify its intended use for screening the potential of substances to act as oestrogen agonists and antagonists *in vivo*

Member 7

The study design of the Uterotrophic Bioassay did not assess the performance sufficiently. The defined data interpretation procedure, prediction model, and a final trial plan, which includes the objectives of the validation study, the study design as well as the data analysis, are lacking.

The number of test substances was not selected adequately to address the performance of the assay. For example: only one antagonist was tested at one time point, so solely the feasibility of the assay to screen for antagonists was shown, its reproducibility and predictive capacity cannot be assessed. Neither the specificity was adequately considered, since only one negative chemical was tested (that, according to literature, has been shown to act as a weak oestrogenic agonist). Before the uterotrophic assay could be used for regulatory purposes, its role in a testing strategy combined with other in vitro and in vivo tests should be defined.

Responses for 9th February, 2004, teleconference

Has 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?*

Member 8

Yes. The test is detecting specifically and with a good sensitivity the tested estrogen agonists. Antagonists, especially weak antagonists were not tested, although their biological relevance is questionable.

Overall, the in vivo assay measuring rat uterine weight changes is well characterized, reproducible in different labs and may be validated.

Observer 1

No, the data derived from the validation study indicate the uterotrophic assay *is not suitable* for use as a screen for the identification of weak uterotrophic agonists. The data from the validation study *are insufficient* to make any conclusion regarding the utility of the test for the identification of uterotrophic antagonists.

There are several procedural and scientific issues that lead to this conclusion. The following comments do not address the procedural issues except to note the following: there are several instances where the designers of this study did not follow the guidance given in OECD guideline on the design and implementation of a toxicity test validation study. The decision not to follow OECD's guidance has decreased the clarity of results derived from the study and made assessment of the study's results difficult and contentious. It is recommended that these issues be fully documented by the PRP so that future validation studies conducted by the VMG incorporate state of the art validation principles.

The scientific issues that prevent one from drawing definitive conclusions on the utility of the uterotrophic assay derive from two fundamental flaws in the design of the validation study, incomplete and inappropriate statistical analysis of the study's results, and unacceptably high between laboratory reproducibility of the test.

Design flaws: There were two fundamental flaws in the design of this validation study. Firstly, an insufficient number of negative test substances were included in the reference set of chemicals. When evaluating the performance of a toxicity test the key performance characteristic that needs to be assessed is whether or not a test has the capacity to correctly distinguish between chemicals that are positive versus those that are negative. In order to assess this performance metric it is necessary to include a sufficient number of positive *and* negative test substances in the reference set of test chemicals. Recent contributions to the toxicity test validation and epidemiology literature document the importance of this measurement and demonstrate why it is necessary to include negative test substances in a validation study (4-6). There are two solutions to this problem. The first is to conduct additional testing of negative test substances and use the results to determine the sum of the sensitivity and specificity of the test. The second is to utilise Monte Carlo simulations to provide estimates of the predictive capacity of the test. The advantage of the second suggestion is that useful results could be obtained quickly without the need for additional animal tests but it would require general agreement that the between laboratory reproducibility of the DBP derived from this study is representative of results from all negative chemicals.

The second design flaw was that the dosing schedule was specified to the participating laboratories. The specification of the dosing schedule controlled a significant source of between laboratory variability. This means that the results derived from the study underestimate the true between laboratory variability to be expected from the test.

The statistical analysis is insufficient to characterise the predictive capacity of the test: There are two problems associated with the statistical analysis of results from the study. The first is that the analysis presented to the PRP is incomplete. The study statistician focused his work on an assessment of the capacity of the test to identify positive test substances. This assessment alone is not sufficient to characterise the predictive capacity of a toxicity test. There are two other assessments needed in order to characterise the test's performance. The first, as noted above, is a measurement of the capacity of the test to correctly distinguish between positive chemicals versus those that are negative. The aim of this analysis is to determine whether the sum of the sensitivity and specificity is greater than one. The importance of this measurement is reviewed in the documents provided to the PRP and in the validation literature (2-4). The inclusion of negative test substances in the study design would have made it possible to determine this metric. The second analysis missing is a quantitative assessment of the effects of the between laboratory variability on the predictive capacity of the test. There are several ways to approach such an assessment including the use of Monte Carlo simulations. An example of such an analysis has been presented to the team. This analysis is critical because it will provide potential users an indication of the reliability of test results.

Has 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?*

Observer 1 – (continued)

No, the data derived from the validation study indicate the uterotrophic assay *is not suitable* for use as a screen for the identification of weak uterotrophic agonists. The data from the validation study *are insufficient* to make any conclusion regarding the utility of the test for the identification of uterotrophic antagonists.

There are several procedural and scientific issues that lead to this conclusion. The following comments do not address the procedural issues except to note the following: there are several instances where the designers of this study did not follow the guidance given in OECD guideline on the design and implementation of a toxicity test validation study. The decision not to follow OECD's guidance has decreased the clarity of results derived from the study and made assessment of the study's results difficult and contentious. It is recommended that these issues be fully documented by the PRP so that future validation studies conducted by the VMG incorporate state of the art validation principles.

The scientific issues that prevent one from drawing definitive conclusions on the utility of the uterotrophic assay derive from two fundamental flaws in the design of the validation study, incomplete and inappropriate statistical analysis of the study's results, and unacceptably high between laboratory reproducibility of the test.

Design flaws: There were two fundamental flaws in the design of this validation study. Firstly, an insufficient number of negative test substances were included in the reference set of chemicals. When evaluating the performance of a toxicity test the key performance characteristic that needs to be assessed is whether or not a test has the capacity to correctly distinguish between chemicals that are positive versus those that are negative. In order to assess this performance metric it is necessary to include a sufficient number of positive *and* negative test substances in the reference set of test chemicals. Recent contributions to the toxicity test validation and epidemiology literature document the importance of this measurement and demonstrate why it is necessary to include negative test substances in a validation study (4-6). There are two solutions to this problem. The first is to conduct additional testing of negative test substances and use the results to determine the sum of the sensitivity and specificity of the test. The second is to utilise Monte Carlo simulations to provide estimates of the predictive capacity of the test. The advantage of the second suggestion is that useful results could be obtained quickly without the need for additional animal tests but it would require general agreement that the between laboratory reproducibility of the DBP derived from this study is representative of results from all negative chemicals.

The second design flaw was that the dosing schedule was specified to the participating laboratories. The specification of the dosing schedule controlled a significant source of between laboratory variability. This means that the results derived from the study underestimate the true between laboratory variability to be expected from the test.

The statistical analysis is insufficient to characterise the predictive capacity of the test: There are two problems associated with the statistical analysis of results from the study. The first is that the analysis presented to the PRP is incomplete. The study statistician focused his work on an assessment of the capacity of the test to identify positive test substances. This assessment alone is not sufficient to characterise the predictive capacity of a toxicity test. There are two other assessments needed in order to characterise the test's performance. The first, as noted above, is a measurement of the capacity of the test to correctly distinguish between positive chemicals versus those that are negative. The aim of this analysis is to determine whether the sum of the sensitivity and specificity is greater than one. The importance of this measurement is reviewed in the documents provided to the PRP and in the validation literature (2-4). The inclusion of negative test substances in the study design would have made it possible to determine this metric. The second analysis missing is a quantitative assessment of the effects of the between laboratory variability on the predictive capacity of the test. There are several ways to approach such an assessment including the use of Monte Carlo simulations. An example of such an analysis has been presented to the team. This analysis is critical because it will provide potential users an indication of the reliability of test results.

The second issue regarding the statistical analysis of results from this study is whether or not the use of analysis of covariance (ANCOVA) is valid. The study statistician's response to my earlier query does not satisfactorily address the concerns raised. The appropriateness of the use of the ANCOVA procedure is questionable and its use requires further evaluation.

Has 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?*

Observer 1 (continued)

Between laboratory reproducibility is unacceptably high: The between laboratory variability in the uterus weight change is high, ranging into the hundreds and even thousands of percent. CVs at this level have a significant negative effect on the predicative capacity of a toxicity test (1). The results from Monte Carlo simulatons shared with the team provide useful quantitative perspective on the consequences of this variability on the performance of the test. Given the between laboratory variability observed in the study it is expected that negative chemicals will result in significant changes in uterus weight in approximately 40% of tests. Furthermore, the simulations estimate that the sensitivity and specificity of the of the uterotrophic assay will be approximately 50% and 79%, respectively, assuming a between laboratory variability consistent with the results derived from protocol B, a cut-off of 15 mg uterus weight change, and a true prevalence of 10%. The 95% prediction interval for a uterus weight change of 20mg ranges between -9 mg to 66 mg! Performance at this level would not serve the regulatory community's need to protect public health. Performance at this level would also result in unnecessary loss of resources due to false positive results. The Monte Carlo analysis provided to the team should be extended to the results from all protocols evaluated in the study. The analysis should also be peer-reviewed to ensure that there is agreement with the approach taken and the conclusions derived from it.

Conclusions: The results from this validation study indicate the uterotrophic assay is not yet ready for use as a general screen for the identification of uterotrophic agonists and antagonists. Since the results from this validation study represent performance under optimised conditions it can be expected that the performance of the test will be worse when put into general use. Releasing this test general with the performance characteristics observed in this study would not provide the public health protection sought by the regulatory community.

The results from this study should be considered a useful pre-validation exercise that provides plenty of information useful for further development of the test and refinement of the protocol. In order to move forward it will be important to identify the sources of the high between laboratory variability and implement steps to control it. When this work is done the test should be reassessed in an appropriately designed validation study.

Responses for 9th February, 2004, teleconference

Summary Conclusions
Based on the information provided in the Submission Package:
a. Does this method adequately identify the potential for test substances to act *in vivo* as possible oestrogen agonists and antagonists?
Member 1 'Probably yes' for agonists, and 'uncertain' for antagonists.
Member 2 This judgment is limited again by the relatively low number of compounds tested. For the compounds that were tested, the assay gave favourable results, which is promising. But a definitive statement can only be made after more elaborate testing, with emphasis on antagonists and negative compounds.
Member 3 Possible oestrogen agonists – yes, but there is also a currently unmet need to be able to determine whether the significant increase in uterine weight observed is due to a chemical acting predominantly via an oestrogenic or androgenic (or other?) effect. Other relevant issues raised during the peer review also need to be addressed satisfactorily (e.g. additional details to be included in a final test protocol). Possible oestrogen antagonists – theoretically yes, but the data available currently are too limited to reach a conclusion.
Member 4 Yes, the Uterotrophic bioassay is considered as a robust and rapid *in vivo* screening assay for possible oestrogen agonists/antagonists, based on the responsiveness in oestrogen sensitive tissue, however, the assay may have some limitations in its sensitivity and specificity compared to other endpoints used to determine a possible oestrogen activity e.g. epithelial cell height, gland number, uterine cell proliferation, lactoferrin protein induction, and measurement of the expression of oestrogen regulated genes or proteins. It is important to be aware of the limitation of the assay in the evaluation of the results. One limitation is that ERα is expressed in the uterus at significant levels, whereas ERβ is expressed in the uterus in moderate levels, limiting the detection of substances binding to the ERβ.
Member 5 Yes, it is adequate but testing of additional antagonist compounds would have strengthened the antagonist issue.
Member 6 All published studies with the Uterotrophic Bioassay show that it is adequate to identify the potential of test substances to elicit the hormonal responses of interest, *i.e.* to act as oestrogen agonists and antagonists *in vivo.*
Member 7 This model could identify compounds that have a biological effect *in vivo* increasing the uterus weight but, a positive result in the uterotrophic assay cannot exclusively result because the tested chemical is an oestrogen agonist since other toxicological pathways can also lead to a stimulation of the uterus growth. On page 87-89 of the background review document the authors describe that a positive uterotrophic result can also occur with non-oestrogens, e.g. androgens, progestins, and growth factors. A definitive conclusion can only be drawn in combination with additional tests such as receptor binding tests. Furthermore, the advantages of the uterotrophic tests have only partially been addressed in the validation study (metabolism of *non-active parental compounds* and toxicokinetics). The validation study is not confirming the ability of the uterotrophic test also to detect active metabolites from non or weak oestrogenic chemicals. In order to decrease the level of false positives the authors suggest performing precursor assays such as ER binding. It is questionable whether an Uterotrophic test is still necessary if the chemicals have demonstrated to be positive *in vitro* assays. As discussed before the metabolic competence for this approach has not been proven. In addition, a positive in vitro result should not automatically lead to an additional *in vivo* experiment. This approach is therefore questionable.
Member 8 See my first comment, for agonists with very good efficiency, for antagonists only one case proved.
Observer 1 No. See comments above. **[to first question – Secretariat]**

b. Are there currently available methods that can provide equivalent or better *in vivo* predictions of oestrogen agonists and antagonist (reliability), with equivalent or better relevance, with equivalent or lower costs, and with equivalent or lower use animals?

Member 1

Having been out of the mainstream of this field for a few years, this reviewer is unaware of a test that does a better job of integrating a biological response. Other tests encompass smaller pieces of estrogen biology; one attraction of the U.B. is that it is an apical, holistic test for showing estrogenicity (and probably some other hormone actions as well). I know of no other tests that have been subjected to any degree of cross-lab validation, so this test is several steps ahead in that area.

Member 2

It is difficult if not impossible to give a judgment here as one compares methods which are in very different stages of development and validation. This validation has shown that the uterotrophic assay is reproducible between labs, its predictivity is still open to questions about specificity and sensitivity. Other assays have not been validated to this extent, and the first impression is that the Nacif assay is more laborious and requires more specific technical skills.

Member 3

None that, to my knowledge, have been evaluated to the same extent as the rodent uterotrophic bioassay (i.e. subjected to a detailed, systematic review of experimental procedures and data integrity and interpretation).

Member 4

The Uterotrophic bioassay is a robust and cost-effective *in vivo* screening assay for oestrogen agonists/antagonists.

The measurement of expression of oestrogen regulated genes is a sensitive and specific approach, however, such methods needs a high level of technical skill, and are not cost-effective.

Measurement of the luminal epithelial cell height or uterine cell proliferation by BrdU labelling or PCNA are alternative cost-effective methods that could be used widely as the Uterotrophic bioassay, however, these methods have not evaluated at the same level as the Uterotrophic bioassay.

Member 5

No, but I think we should remain alert to the possibility of other assays being developed. This is a dynamic field and the proposed uterotrophic assay the best at this time but it could be supplanted in the future.

Member 6

Recent studies of gene expression (analysis by conventional or microarray techniques; e.g. Diel et al. 2000; Newbold et al. 2001; Naciff et al. 2003), along with an analysis of classical gravimetric and histological parameters in target tissues, have indicated that changes in gene expression provide useful additional information, and can increase the sensitivity for detecting responses to oestrogens.

However, such methods presently cannot compete with regard to costs (exceed by far those of the classical uterotrophic response assay), and do not result in lower use of animals. Furthermore, as yet they have not been shown to provide better or more reliable *in vivo* predictions of oestrogen agonists and antagonist.

Member 7

No other *in vitro*/*in vivo* method has been validated to reassure the reliability or relevance to predict hormonal activity yet. However, a validation of *in vitro* assays should be considered as a priority, these will have lower costs, and fundamentally no use of animals (3R's). In fact in this validation exercise, a total of 7200 animals were used. The study did not take animal welfare into account i.e. in accepted validation criteria 3-4 labs are sufficient for a formal validation study.

Member 8

No, especially not for mammalians. The inter-species variability of the estrogen function is high. It is true that in case of xenoestrogen contamination also other species should be checked.

Observer 1

No comment.

Responses for 9th February, 2004, teleconference

c. Discuss conditions/limitations/restrictions that may affect the intended use of Uterotrophic Bioassay, and that are justified based upon the current presence or lack of scientific evidence.

Member 1

Again, uncertainties about the effects of testosterone (which should be caught in a receptor-binding assay), and the consistency of the ability to identify antagonists. Solubility might also be an issue, but that would also tend to reduce exposure to most other animal systems. A high background weight, due to diet or bedding, or a malicious Fate, would limit the usefulness of an assay, but this should be intermittent and not an inherent structural flaw in the biology of the assay itself.

Member 2

Limitations are secondary to the low number of compounds tested, with special reference to antagonists and negative compounds. We have insufficient information from this validation study about sensitivity and specificity of the assay. Still, in a tiered approach as a follow-up to receptor-specific in vitro assays, the test can play a useful role as the first in vivo screen for estrogenicity performed.

Member 3

Lack of clarity on the context / scope of use of the bioassay, and still a rather limited understanding of the assay limitations (need for a larger database) and data interpretation for risk assessment purposes, are the main concerns. I don't think the uterotrophic bioassay is ready to be proposed for routine regulatory use for testing chemicals until there is a clear rationale for how it would be used in a tiered testing strategy and how the resulting data would be used in regulatory decision-making. Based on the scientific evidence submitted for review, the bioassay appears useful for in-house decision-making in an overall scientific weight-of-evidence type of approach (relatively flexible) but not yet at the stage where it should be incorporated into a more-rigid regulatory testing framework.

Member 4

Due to the appearance of false negative and positive with the Uterotrophic bioassay, the need for clear criteria for data acceptance is necessary, e.g. maxima for acceptable vehicle control uterine weights, and how to interpret a modest increase in uterine weight (20 to 40 % increase).

The selection of doses used in the Uterotrophic bioassay may also influence the outcome of the test if mortality or systemic toxicity is achieved at the doses tested.

The lab experience/technical skill may also influence the outcome of the Uterotrophic bioassay.

The EU policy related to toxicological testing is to reduce the number of animals used in toxicological testing, how this will influence the use of the Uterotrophic bioassay as an *in vivo* screening test is uncertain.

Member 5

Over-all, I think the uterotrophic assay has been adequately described and the consistency among the test laboratories was quite impressive given that there were variables uncontrolled such as food, caging, and strain of rat used. I am concerned about the \possible confound of phytoestrogens and, in future tests, some control over the upper concentration of phytoestrogens would be warranted.

Member 6

All published studies (conducted within and outside the OECD Program) with the Uterotrophic Bioassay justify its intended use for screening the potential of substances to act as oestrogen agonists and antagonists *in vivo*.

One (minor) uncertainty pertains to the question how sensitive the assay will detect <u>weak</u> oestrogen antagonists. But, this is not based on principle doubts, just a poor data basis of studies demonstrating such effects for agents other than the prototypical antioestrogens.

Member 7

The Uterotrophic bioassay aims to detect strong and weak oestrogens and antioestrogens. Since very high concentrations (up to 1000mg/kg) have to be applied in order to receive some biological responses it is questionable if these results are relevant for human hazard assessment.

The protocols for detecting oestrogens remain to be optimised, particularly for minimising all sources of variation to make the protocols more reliable and economical with regard to use of animals.

Responses for 9th February, 2004, teleconference

c. Discuss conditions/limitations/restrictions that may affect the intended use of Uterotrophic Bioassay, and that are justified based upon the current presence or lack of scientific evidence.
Member 7 (continued) It is impossible to fully assess the predictivity of the assay in terms of its specificity and its sensitivity, without the use of more negative controls. An additional study to analyse the reliability and relevance of the assay for detecting anti-oestrogens is necessary to evaluate the predictive power. The design and outcome of this validation study do not conform with the validation criteria as discussed in the OECD document No 34 that is currently under discussion for new and revised tests. Although it is claimed several times that the study followed international principles of validation, several aspects of validation principles have been ignored e.g. study design, selection of laboratories, selection of test chemicals etc. The helpful separation of pre-validation, including protocol transfer, protocol refinement, as well as a small-scale study, and the formal validation process was followed insufficiently resulting in mixture of these tasks in three different study parts. The overload with reports makes it difficult to extract the crucial information needed to peer-review the validation.
Member 8 The estrogenic response detected by the uterotrophic assay had been demonstrated not only by the OECD validation procedure but also by many other laboratories. The scientific evidence is enough to use it for in vivo screening with a well- and clearly-documented protocol.
Observer 1 The results from the validation study indicate there is unacceptably high between laboratory variability in results. Monte Carlo simulations suggest that the test will not reliably distinguish weak uterotrophic chemicals from those that have no uterotrophic activity. Additional work should be undertaken characterise the consequences of high between laboratory variability on the predictive capacity of the test.

References from Bruner

1. Bruner LH et al. Validation of alternative methods for toxicity testing. TOXICOLOGY IN VITRO 10[4], 479-501. 1996.
2. Bruner LH et al. An investigation of new toxicity test method performance in validation studies: 1. Toxicity test methods which have predictive capacity no greater than chance. Hum.Exp.Toxicol. 21[305], 312. 2002.
3. Bruner LH et al. An investigation of new toxicity test method performance in validation studies: 2. Comparative assessment of three measures of toxicity test performance. Hum.Exp.Toxicol. 21, 313-323. 2002.
4. Bruner LH et al. An investigation of new toxicity test method performance in validation studies: 3. Sensitivity and specificity are not independent of prevalence or distiribution of toxicity. Hum.Exp.Toxicol. 21, 325-334. 2002.
5. Choi BC. Causal modeling to estimate sensitivity and specificity of a test when prevalence changes. Epidemiology 1997;8:80-6.
6. Guggenmoos-Holzmann I, van Houwelingen H. The (in)validity of sensitivity and specificity. Statistics In Medicine 2000;19:1783-92.