Fish Short-term Reproduction Assay Peer Review – EPA Response

Peer reviewer comments are organized by OECD GD34 validation criteria with some added categories associated with more technical comments and recommendations from the reviewers. The reviewers comments are attributed as BEB = Bengt-Erik Bengtsson, Richard DiGiulio = RDG, Deborah MacLatchy = DM, Mark McMaster = MM, and Helmut Segner =HS.

Overall, it is concluded that that the Fish Short-term Reproduction Assay is valid for its intended purpose. However, some additional clarification and details are needed in the test method protocol.

EPA thanks the reviewers for their assistance in reviewing the voluminous validation materials and is very appreciative of their helpful and constructive comments.

Comment	EPA Response
General comments: **RDG/** The Fish Short-Term Reproductive Assay described in the Integrated Summary Report (ISR) and supported by ample supporting materials is overall an excellent approach for Tier 1 screening for chemicals that might perturb the hypothalamic-pituitary-gonadal (HPG) axis. The rationale for the design of the 21 day test, the endpoints selected, and approaches for analyzing data clearly have been carefully thought through, and are very well described is this ISR. Chemicals selected for developing and optimizing the assay are diverse and cover major known mechanisms of HGP interference, as well as chemicals whose mechanisms are unknown and chemicals not thought to interfere with the HPG (i.e., negative controls). Thus, the studies and data supporting this assay are very impressive, and lead to the conclusion, by this reviewer at least, that this assay will be highly effective for its intended purpose. Responses to specific charge questions are provided below. **MM/** Overall, I thought that the Integrated Summary Report for the Fish Short-Term Reproduction Assay was quite detailed and informative. I was fairly disappointed with the quality of some of the interlaboratory studies although I acknowledge that performing the assay is quite difficult. I would recommend that more work is conducted on ensuring that fish as similar as possible in fecundity per day are used to start the assay. This was not done very well by some of the labs in the interlaboratory studies. The following is my response to the Charge Questions. **HS/** As a reviewer, my primary task is not to cheer but to critically evaluate the fish assay. Having said so, however, I would like to emphasize that my impression of the assay and its utility is highly positive, and that the development and validation work presented in the ISR is indeed most impressive.	The positive comments on the Assay are acknowledged and appreciated.
1) The rationale for the test method should be available. **BEB/** Yes, the purpose is well described and all necessary details and explanations are included and well written. In particular, the presentations in *att-f* and *att-g* are very useful in this respect.	In general, the reviewers considered the rationale for the test method to be well described. Some suggestions were provided for

RDG/ Yes. The purpose of providing a screen for the detection of chemicals that may perturb the hypothalamic-pituitary-gonadal (HPG) axis, particularly through (anti-) estrogenic and (anti-) androgenic effects, is clearly described. For readers less familiar with Tier 1 and 2 terminology and testing, it would be helpful to describe more fully what questions are addressed and methods likely employed for Tier 2 testing. Also, it would have been helpful to place the key statement that nay significant effect on one or more core endpoints of this assay triggers Tier 2 testing (p. 81, section 6.2, Data Interpretation) much earlier in the document, and to justify it.	additional clarification and these will be considered in the final revision of the test method. In conclusion, this criterion is judged to be met.
DM/ The purpose of the fish short-term reproduction assay with fathead minnow, to provide a Tier 1 *in vivo* assay to detect chemicals that alter fish reproduction, morphology and certain biochemical endpoints reflective of hypothalamo-pituitary-gonadal (HPG) axis functioning, is clearly stated. That the bioassay is a "first screen", intended to determine whether further testing is necessary, is clear, as is the point that it is not intended to specify mechanism of action.	Additional information on the Tier 2 test methodology can be found on the public website for the EDSP (www.epa.gov/scipoly/oscpendo). A significant effect on a core endpoint in the fish assay would lead to an interpretation that there is a potential endocrine interaction. Tier 2 testing is triggered only after a weight-of-evidence review of the entire EDSP Tier 1 battery of assays.
MM/ The fish short-term reproduction assay with fathead minnows is designed to detect changes in spawning, morphology and specific biochemical endpoints that reflect disturbances in the HPG axis. It is important to recognize that the assay is not intended to quantify or confirm endocrine disruption, or to provide a quantitative assessment of risk, but only to provide suggestive evidence that certain endocrine regulated processes may be sufficiently perturbed to warrant more definitive testing. Although some endpoints may be highly diagnostic (*e.g.*, vitellogenin induction in males and tubercle formation in females), not all endpoints in the assay are intended to unequivocally identify specific cellular mechanisms of action, but collectively the suite of endpoints observed do allow inferences to be made with regard to possible endocrine disturbances and thus provide guidance for further testing. Overall, the stated purpose of the assay is fairly clear. The second sentence in the purpose does help to clarify this, but I am not sure that suggestive evidence is suitable for a regulatory assay. I think a clearer description of the level of change required in specific endpoints is required. Suggestive? Is this significantly different then reference? Dose response relationships within an endpoint? One endpoint or multiple endpoints? Endpoints that are linked in mechanisms? ie. Estrogenic responses, vtg in males, female characteristics? Are these suggestive changes? Then at the end of the purpose it is stated that collectively, the suite of endpoints allow inferences with regard to endocrine disturbances. What is required to determine suggestive evidence? Are all changes in spawning, morphology and biochemical endpoints necessarily a reflection of a disturbance in the HPG axis? If a fish decides not to spawn or release their gametes because the environment is not suitable for the development/hatching and potential survival of their offspring, is this endocrine disruption? Is it a disturbance in the HPG axis or just a response of the axis to this decision? Question? If a fish decides not to release its eggs, it will begin a process to break down its eggs and re-adsorb the energy. Steroid levels will change due to the change in the decision to spawn. These may not all be due to endocrine disruption or alterations specific to the chemical being tested, just a decision to not spawn in an environment not fit for survival of the young. I agree that responses such as these are also of regulatory importance, but may not be a direct effect on the endocrine axis. Support from other Tier 1 assays will help to support the mechanisms of action. What is the result of the Tier 1 screen, if a chemical reduces the number of eggs laid but no binding to receptors occurs, induction of Vtg is not affected etc? Does that result then represent the potential for endocrine disruption? It is a very difficult phenomenon to regulate as almost everything can result in some change in an endocrine function. Is there also some form of recognition as to the potential level for these chemicals to reach in the environment? Almost everything will result in some form of endocrine response if the concentration is high	Agree that the term "suggestive" is overly simplistic. Data interpretation will use weight-of-evidence determination. Statistically significant changes in a core endpoint will indicate potential endocrine interaction, but as previously stated a weight-of-evidence review which includes the entire Tier 1 battery will be performed to determine if one can conclude no concern or that further testing is required. In the example for reproduction impairment, this may or may not be by direct effect on an endocrine axis and a weight-of-evidence review would inform further action (need for additional data).

enough. Low DO results in endocrine stress related responses which in turn alter reproductive steroids due to stress related conditions and a response of the gonads to elevated cortisol levels. EDC's are very difficult to regulate for this reason. **HS/** The purpose of the assay is almost but not 100 % clear. On p. 11 of the ISR, the purpose of the fish assay is described as follows: "The fish short-term reproduction assay with fathead minnow is designed to detect changes in spawning, morphology and specific biochemical endpoints that reflect disturbances in the HPG axis." This formulation would agree with the overall EDSP strategy that the purpose of tier 1 assays the initial sorting and prioritization of chemicals that warrant further testing, but not to confirm any specific mechanism, mode of action or adverse effect. This description of the assay's purpose, however, is not consistently followed in the ISR. For instance, on p. 6 it is said that the Fish Assay is "one of the (anti-)estrogen- and (anti) androgen-relevant screening assays". Or on p. 56 it is said that the Fish Assay "has been optimized primarily to detect estrogen and androgen agonists/antagonists". In the majority of cases, disturbances of the HPG axis may indeed be directly or indirectly caused by chemicals that act as agonists/antagonists of estrogens and androgens, however, this mode of action is not exclusive and other modes of action may lead to disruption of the HPG axis as well. Reproductive endocrinology goes beyond the sex steroids, incorporating, for instance, influences of endocrine systems such as the GH/IGF-I system (e.g., Melamed et al. 1999, Endocrinology 140:1183; Negati et al., 1998, Fish Physiol Biochem 19:13), neuroendocrine systems, or nutritional endocrine factors such as leptin, etc. Thus, in my view, formulations focusing this assay primarily on estrogen- and androgen-signaling pathways as target processes unnecessarily restrict the more general scope of the assay as defined on p. 11. i.e. "to detect changes in spawning, morphology and specific biochemical endpoints that reflect disturbances in the HPG axis" (p. 11). I agree that this is a little bit semantics, still I feel in order to avoid confusion it would be helpful to be as consistent and precise as possible. One addition might be helpful in specifying the purpose of this "reproduction assay": The term "reproduction" can refer to reproductive parameters of adult fish such as fecundity or fertility, but sometimes the use of the term also includes parameters of offspring performance such as hatchability and early life stage survival. Accordingly, as stated on p. 11, disrupted HPG axis functioning may not only lead to impaired adult spawning but also to impaired hatching and larval survival. Actually, the latter two parameters are included in the assay protocol as "optional parameters", but they are not used as "core" parameters. Thus, the reproduction assay is primarily an assay to assess reproductive parameters of adult fish (with the option to be extent to offspring recruitment), and it may be important to point out this fact anywhere in the description of the purpose of the assay.	Agree. The assay is not limited to estrogens and androgens, but is capable of detecting changes in the HPG axis which includes the estrogens and androgens.
2) The relationship between the test method's endpoint(s) and the biological phenomenon of interest should be described. **BEB/** Yes, all relevant aspects are presented. Several of the documents referred to contain very detailed and comprehensive information necessary to describe the test species suitability and the scientific background and relevance of the proposed effect variables. **RDG/** Yes. The overall design and endpoints selected are generally highly appropriate for screening for HPG perturbing chemicals, particularly (anti-) estrogenic and (anti-) androgenic compounds. One concern is for the incorporation of histopathology as a key endpoint in the assay. While clearly a powerful method for discerning chemical effects, and as noted for potentially linking HPG-associated biochemical effects with organismal effects (e.g., fecundity), it may	Agree with the comments made. In conclusion, this criterion that the test endpoints and phenomena of interest are described and is therefore judged to be met.

be overkill for a screening assay. Histopathology, in contrast to the other endpoints, requires very specialized training, and moreover is very labor-intensive. It would seem to kick this assay a major notch up, in terms of costs, time, and expertise required. Moreover, in the protocol optimization and inter-laboratory validation studies described, it appears that significant histopathology changes were accompanied by significant responses in other endpoints (fecundity, morphology and/or biochemistry). Also in several instances in the optimization studies, histopathology results appeared inconsistent (e.g., estradiol, p. 30; bisphenol A, p. 37; 17□-trenbolone; p. 40, flutamide, pp. 47-48). This issue gets back to that raised above, concerning the distinction between Tier 1 and 2 testing. Perhaps histopathological analysis is more appropriate for more in depth Tier 2 testing than Tier 1 screening.	Yes, histopathology will increase costs and time, but is considered essential to reduce false negatives.
Another concern is the aforementioned statement under Data Interpretation (p. 81) that "any significant effect in one or more of the core endpoints of the assay (fecundity, histopathology, GSI, sex steroid measurements, vitellogenin, and secondary sex characteristics) should be considered a positive response…..and supports further testing of the compound in the Tier 2 assays of EDSP." This appears perhaps overly conservative, as at high doses, many (most?) compounds are likely to reduce fecundity and cause tissue damage (including but not limited to gonads). For example, the chemical used as a negative control in the validation studies (sodium dodecyl sulfate, SDS) would clearly meet this trigger criterion (although these results are interpreted by the authors as evidence that SDS may be an endocrine disruptor).	Agree. This has to be taken into consideration as part of a weight-of-evidence determination.
DM/ The fish short-term reproduction assay is biologically and toxicologically relevant to the stated purpose. The endpoints measured in the assay directly correspond to: (a) the endocrine control of reproduction and (b) the biological process of reproduction in this species, including egg production (and can be easily expanded to include measures of fertilization success and early lifestage development). Therefore, this is an appropriate bioassay, using suitable endpoints, to determine if there is potential for impacts to fish reproductive status. The appropriateness of the bioassay can be supported by the following: • Route of exposure: flow-through waterborne exposures in fish bioassays ensure a steady-state of exposure, are environmentally relevant, and can be quantified, standardized and validated; • The test chemicals used for the interlaboratory validation and in other peer-reviewed studies covered a range of endocrine-mediated pathways of effect, including: o Estrogen agonists: 4-t-octylphenol (interlab study); methoxychlor, 4-nonylphenol, bisphenol A, etc. (peer-review studies); o Androgen agonists: 17α-/β-trenbolone (peer-review studies); o Androgen antagonists: prochloraz (interlab study; also a steroid biosynthesis modulator); flutamide (peer-review studies); vinclozolin (interlab study, peer-review studies); o Steroid biosynthesis modulators: fadrozole (peer review studies); and o Multi-modal compounds: ketoconazole, prochloraz (interlab study; peer-review studies). • Overall exposure protocol: has been developed by paying attention to the particulars of the biology of reproduction of this fish species (e.g., spawning ratios, food requirements, photoperiod, temperature, spawning behaviours, spawning substrate requirements, etc.). Importantly, the comparison of group spawning, pair breeding and non-spawning bioassays has been carried out; • Pre-exposure period: reproduction in fish can be highly variable and the use of the pre-exposure period eliminates a degree of variability, thus increasing the power of the test	

during the exposure period. As well, pre- and post-exposure endpoints can be compared, strengthening the statistical analysis; • Exposure period: the 14- to 21-day period allows for 5 to 7 spawning periods of individual female fish to be encompassed within the bioassay (this period has been determined to be adequate to indicate changes by compounds with known mechanisms of action, such as estrogen and androgen agonists); • Endpoints: have been developed to ensure the bioassay is able to directly and indirectly detect effects on the HPG axis o Survival: ensures chronic exposures are being carried out and that the fish stocks used are healthy (controls can be used to measure and maintain interassay standardization); o Behaviour: in addition to general toxicity behaviour, reproductive behaviours are noted. This can be important as effects in behaviour can become apparent before other morphological or survival endpoints are affected; o Fecundity: an important population-level indicator, it integrates effects on the HPG axis, general toxicity, behaviour, etc. and is a consistent response of fish affected in the HPG axis; o Fertilization success: another important integrative endpoint, it is an indirect measure of egg and sperm viability; o Embryonic/larval development: when assessed, provides a measure of the longer-term impacts of parental and on-going exposure during reproduction on offspring development; o Secondary sex characteristics: this is both a direct and indirect method of assessing the potential for particular chemicals to alter reproduction, especially for receptor agonists and antagonists; o Gonad histology: assessment of gonadal histology is important for two reasons: (a) it provides a direct measure of exposure effects (control vs. treatment); and (b) also ensures that endocrine status (e.g., plasma steroid levels) are contextualized by knowing the reproductive status of individuals (i.e., comparing animals in the same or different reproductive stages within and among exposures). The sum total of histological changes generally indicate a pattern of response; o Plasma sex steroids: are an indicator of whole-organism reproductive endocrine status, and therefore directly representative of HPG axis status; and Vitellogenin: levels of plasma (or mRNA) vitellogenin are a direct indicator that a compound is estrogenic (i.e., binds to estrogen receptor and initiates response) and may also indicate compounds with (anti-) estrogenic mechanisms. **MM/** I realize that a battery of tests will be used in the Level 1 screen of chemicals. The short-term fish assay incorporates a number (at least 3) of the other battery endpoints within the whole fish test. What is the requirement of the tests in the batteries to support one another in their results? Does an androgenic response in the androgen receptor assay also require androgenic responses in the short term fish test? I realize that this is a question that will be asked in the next stage of battery development, but it should be indicated here as well to help us determine the relevance of the assay. Biologically speaking the assay is solid in its ability to detect changes in reproduction in this fish species. However, not all responses seen will necessarily reflect a disturbance of the HPG axis directly (binding	The EPA understands the nature of the reviewer's (MM) concerns and questions. However, these questions are relevant to the Tier 1 battery overall which will be considered by the FIFRA Scientific Advisory Panel.

to a receptor, increases in the production of protein etc). If a fish determines that the environment is not suitable for survival of its young, it will not release its eggs. Is this considered an endocrine disturbance? If for example, no eggs are released in exposed fish, this may be the case. I would much prefer a dose response decrease in egg production then a complete cessation of spawning to suggest an endocrine disturbance of the substance. It also incorporates a number of different endpoints of reproductive function and success. This increases the tests biological and toxicological relevance relative to a number of the other Tier 1 battery tests. In fact it is approaching a Tier 2 level test. Toxicologically speaking the test is also very relevant. Unlike some of the other Tier 1 screening assays, this incorporates a whole animal which will take into account, metabolism of the compound, excretion, uptake, etc. It is more toxicologically relevant than most of the other Tier 1 battery. Binding to a receptor does not necessarily result in an effect in a whole animal test which helps support this tests relevance toxicologically speaking.	The apical endpoints in this assay demonstrate the importance of an intact organism for use in determining the potential for endocrine disruption. The integrative nature of the endocrine system (and its overlapping components with other systems) is reflected in the responses observed in apical endpoints within the whole organism, but not always in specific biochemical endpoints.
HS/ A screening assay should rely on parameters that are relevant to the purpose of the assay, in this case, to detect the potential ability of a test compound to disrupt the HPG axis. As outlined in attachment A, the relevance is difficult to discuss if it comes to tests for endocrine disruption. The Fish Assay relies on a combination of apical endpoints (fecundity) with more diagnostic endpoints (histopathology, biochemical parameters). Given the scientific understanding of the action of EDCs in fish, and the limited empirical knowledge on the correlation between the EDC impact and the selected endpoints, the combination of endpoints chosen for the Fish Assay appears to be sound and both biologically and toxicologically relevant. The strengths of the endpoints are appropriately discussed in the ISR (mainly on p. 83 ff), their selection is well defendable and further supported by the information provided in the DRP (attachment B). Of course, one may think of additional, more mechanistic endpoints (see, for instance, the study of Villeneuve et al., 2007, Tox Sci , 98:395), however, one has to keep the Fish Assay practical for routine testing. Further, since the Fish Assay does not aim to reveal specific mechanisms, it is questionable whether more detailed and specific endpoints would substantially improve the relevance of the assay. Do the results from the validation tests support the biological and toxicological relevance of the endpoints used in the Fish Assay? The answer to this question is clearly positive if it comes to compounds acting through the steroid receptors or interfering with steroid synthesis, however, it gets more equivocal for substances with less clearly defined or multiple actions. The hallmark in the assay response to estrogen receptor agonists such as octylphenol is the increase of VTG in males. This response would trigger tier 2 testing. In the interlaboratory validation, this was found by all three laboratories, indicating that the effect is rather robust. The VTG response is accompanied by reduced androgen levels, impaired testicular maturation and loss of secondary sex characteristics in males (overall a typical feminization response), and by reduced fecundity together with increased oocyte atresia in females. Treatment with androgen receptor agonists such as 17beta-trenbolone leads to clear changes of secondary sex characteristics in females, accompanied by reduced fecundity and circulating E2 levels. These effects have been consistently reported from various short-term reproduction assays with FHM, and they would trigger the substance for tier 2 testing. The database for estrogen receptor antagonists such as tamoxifen is rather small, however, the typical response in the Fish Assay seems to be reduced fecundity together with reduced VTG. Again, this would be sufficient to identify the compound for tier 2 testing. Fecundity together with secondary sex characteristics seem to be the most consistent endpoints responding to androgen receptor antagonists such as flutamide or vinclozolin. In the interlaboratory study, all three laboratories observed reduced fecundity and tubercles at least for the highest test concentration of vinclozolin, while one laboratory detected the effects even at medium and low vinclozolin concentrations. While the reduced fecundty indicates that the	The text by HS is largely accurate. However, in his discussion of prochloraz he restricts his analysis solely to the ring test results where one laboratory did not see a decrease in VTG. However, the fungicide consistently decreases VTG in every study in the Duluth EPA laboratory (e.g., Ankley et al. 2005), and in the OECD Phase 1B studies. So, from a weight-of-evidence perspective prochloraz is easily flagged as an inhibitor of steroidogenesis (with aromatase as one of its likely targets). Based on assay results, ketoconazole should be considered an inhibitor of steroid production. An effect in gonad histopathology (proliferation of interstitial cells in the testis) was consistently demonstrated in the Duluth EPA laboratory (Ankley et al. 2007) and in the ring test which appears very specific to the disruption of endocrine function. We have never seen this sort of "compensatory" response in studies with other chemicals that might be considered general stressors. In fact, the response observed is quite analogous to thyroid cell proliferation in frogs exposed to thyroid antagonists.

compounds are reproductive toxicants, the additionally affected sex characteristics point to an endocrine etiology and therefore would qualify the test compounds for tier 2 testing. The only response that has been consistently reported for exposure of reproductively active FHM to steroid metabolism modulators such as fadrozole is the reduction of VTG in females. This was also confirmed in an interlaboratory study (OECD 205). An interesting compound in this context is prochloraz which is an aromatase inhibitor but apparently can also act as estrogen- or androgen antagonist. In the interlaboratory study, prochloraz led, in contrast to fadrozole, not to a consistent decline of female VTG, since only two out of three laboratories observed this effect. However, all three laboratories found altered fecundity and male gonad histopathology. Similarly, ketoconazole, a substance with multimodal action consistently altered male GSI and gonad histopathology but had no consistent effect on fecundity. Nevertheless, with these findings neither prochloraz nor ketoconazole would have been falsely classified to be negative but both would have been subjected to tier 2 testing. In conclusion, for the aforementioned substances, the Fish Assay would have correctly classified the compounds as compounds with suspected interference to the HPG axis which need to be further tested tier 2. Based on this, it appears that the assay and the selected endpoints are suitable for the stated purpose, i.e. to detect changes in spawning, morphology and specific biochemical endpoints that reflect disturbances in the HPG axis. The relative sensitivity of the various endpoints is more or less in line with the idea of the "biological hierarchy", i.e. the biochemical and histological endpoints tend to be (slightly) more sensitive than the apical endpoint, fecundity. Less conclusive than the assay results on positive substances are the assay results on the negative compounds (perchlorate, permanganate, SDS) or on compounds with unclear action (atrazin, PFOS, prometon) (see below). It remains to be shown if this is a matter of "bad luck" or a principal problem. One word on the species selection: Although myself working with the zebrafish, I have to admit the advantages of FHM as a test species: gonochoristic development (more relevant in the life cycle than in the adult reproduction test), external sexual dimorphism, relatively large size (ease of plasma collection !), relatively good toxicological database. Unfortunately – in my view - , the Fish Assay protocol does not make use of another advantage of FHM, that is the possibility to use a pair-breeding protocol (see below). A screening assay should yield data that can be interpreted as either negative or positive for determining the necessity to conduct tier 2 tests. This requirement includes two questions: first, what are the criteria to decide whether a compound is positively judged as potential disruptor of the HPG axis, and second, what is the risk of false positives and negatives to be generated by the assay ? Well-defined criteria to decide to decide whether or not a test compound is to be classified as suspected EDC and should be subjected to tier 2 testing – under full avoidance of false negatives and partial avoidance of false positives – are crucial for the success of the Fish Assay. The ISR discusses provides no discussion on this aspect, but states on p. 81 that "any significant effect in <u>one or more</u> of the <u>core endpoints</u> of this assay (fecundity, histopathology, GSI, sex steroid measurements, vitellogenin, and secondary sex characteristics) should be considered a positive response in the Fish Short-Term Reproduction Assay, and supports further testing of the compound in the tier 2 assays of EDSP". In using this decision criterion, none of the substances tested within the Fish Assay validation work would have been classified to be negative. Even the thyroid-disrupting perchlorate, which was intended as negative compound (p. 59 ff), would have to be classified as a suspected disruptor of the HPG axis and subjected to tier 2 testing, since it leads to significant changes in gonad histopathology. Actually, with the current selection of endpoints the Fish Assay appears to be more at risk for false positives than for false negatives. For instance, histopathology may help to avoid false negatives, as outlined on p. 62 of the ISR, but at the same time it may increase the number of false positives, since with the still rather limited knowledge on gonad histopathology of fish, we are	Correct, this is why histopathology is considered such an important endpoint in the assay.

often not able to discriminate between gonad changes due to an endocrine mode of action and gonad changes due to other modes of toxicity. I wonder whether the approach suggested in the ISR (a significant change in onde endpoint is sufficient for a positive classification) may lead to too many false positives. For instance, fecundity is an endpoint that, as stated in the ISR on p. 82, is also responsive to non-endocrine stresses. Provided that a test agent alters only fecundity, but none of the other assay endpoints, it would have to be classified as potential HPG axis disruptor, although a fecundity change alone – in my view – qualifies a compound as reproductive toxicant but not necessarily as a compound with endocrine activity. Thus, the reliance on one single endpoint to classify a compound as potential EDC that has to be submitted to tier 2 testing may lead to false positives. The question is if we would leave this "one endpoint" approach and would require that at least two endpoints, perhaps even a combination of an apical with a diagnostic endpoint, are necessary to decide whether a test agent is suspected as endcorine-active, would this lead to false negatives ? Among the substances tested in the validation process, ketoconazole and SDS could have been critical cases if an apical/diagnostic combination would have been required (see tables 5.3a and b), but not if just any combination of two endpoints would have been required. The compound used as negative control, perchlorate, however, would have been classified indeed to be negative, since only one endpoint (gonad histopathology) showed a significant change. Currently, the available database is too small to provide a conclusive answer on the most efficient and reliable approach for the positive/negative decision. Therefore, for the time being, the more conservative approach as suggested in the IRS to discriminate between positive and negative test compounds, – i.e. a significant change in just one endpoint is sufficient – appears to be wise from the precautionary point of view.	
3) A detailed protocol for the test method should be available. **BEB/** Generally: Yes! However, the problem of miss-sexed fish at start of the experiment is a specific problem with fathead minnows that I have experienced personally and it is also described as a serious problem in *att-d.* The problem is that some males, due to e.g. the suppression by dominating males among stocked fish, may lack secondary sex characteristics and therefore may be wrongly taken for females. This will cause a sex ratio different from the intended and may also affect the results and the statistical power negatively. However, a critical variable is the histopathology, which is also emphasized in the ref. *att-b*. It is very important that there will be histopathological training activities, i.e. specified in the fathead minnow histological endpoints of the test as presented in OECD 2004 (ref. No. 277 in the above document) and in particular in the excellently useful document *att-h*, associated with authorization of the test laboratories. As pointed out in ref. *att-c* the inter-laboratory differences in e.g. VTG and hormone analysis need to be minimized (round robin exercises and benchmark data may be helpful). Similarly, inter-laboratory variation in results with ELISA test kits also has to be minimized. These are general problems and are not specific for this particular assay. The experimental set-up, number of test concentrations, males/females, replicates and description of the statistical background and alternative methods is good. However, it is preferable that the number of statistical methods are narrowed to a few and declared mandatory, i.e. the same statistical test package should be applied by everybody. The borderline between the use of non-parametric vs. parametric test should be clear-cut. Excellent result protocols are presented in *att-g* and will yield sufficient and clear documentation of test conditions and results. **RDG/** The answer to this question is generally 'yes." Certainly the objective of the assay is clearly laid out. In most cases, adequate information for conducting specific assays is provided, and/or primary	Generally agreed. This criterion that a detailed protocol be available is judged to be conditionally met. Recommendations for protocol improvement have been made by the peer reviewers and these will be used in revising a final protocol. It is expected that improvements in commercial ELISA kits with specific FHM antibodies will improve the reliability of these kits. [refer to Jensen and Ankley 2006] As suggested by BEB a standard data reporting spreadsheet and statistical program (or SAS macro) which was developed for the interlaboratory study will be made available to facilitate routine statistical analysis for the assay.

references describing assays are cited. Some clarifications would be useful however. The technique for blood collection is not entirely clear, and the figure in the supporting document (Attachment G, page 57, Figure 6) is of too poor quality to be helpful. Perhaps a web link to a brief video demonstrating the technique would be helpful. For the vitellogenin (VTG) assay, two techniques for measuring the protein are discussed, ELISA and RIA, as well as the alternative of measuring the messenger RNA (mRNA) that codes for VTG (pp. 27-28). While not making a specific recommendation and allowing a laboratory to make its own decision of which assay to use may have merit, it would seem helpful and appropriate for such a recommendation to be made. This would be particularly helpful for laboratories less familiar with various options for a required endpoint in a Tier 1 assay, such as VTG. While the ELISA technique does appear favored, a clearer endorsement of it seems appropriate. Similarly, a more direct recommendation for sex steroid assays would be helpful (p. 27), and a recommendation for a specific dose of MS-222 for anesthesia rather than a range (stated as 100-250 mg/L, p. 26). Presumably, the lowest concentration that is effective (apparently 100 mg/L) would be preferred in order to minimize any possible side effects on endpoints measured. However, if there are reasons for using different concentrations within this range for this assay, they should be stated. A related issue is that of the pH range suggested for conducting the say, which is 6.8-8.3 (p. 18, Table 3-1). This seems a rather broad range considering that many chemicals of interest are likely to be weak acids or bases, for which variables such as ionization and water/lipid solubility are highly pH-dependent. A narrower pH range would seem preferable, or at least some discussion of the rationale for this stated range. Approaches for compiling data and performing statistical analyses are reasonably well-discussed (pp. 65-66), although again are some ambiguities concerning selections among different options. For example, it is at times stated, in regards to a statistical approach, that "it has been recommended" followed by a reference (e.g., Battelle 2006). It is not clear why it isn't simply stated "it is recommended…." There is no discussion of the format/outline for the final report for a Tier 1 screening assay; this may be helpful.	Alternative presentations of the blood collecting technique will be considered. The original image is included in the protocol. [This comment refers to the scanned copy of EPA 2002 provided reviewers of poorer quality than the original. The protocol does contain the original color photo.] The final protocol will recommend ELISA methods for plasma VTG analysis. The protocol provides specific concentration; this comment refers to EPA 2002. The lowest concentration, as indicated in the next sentence, is the recommended one. A range of concentrations was provided in the ISR as background information from the range of studies completed. Agreed, additional guidance on selecting an appropriate pH within the range suggested will be made to the protocol.
DM/ The objectives of the bioassay are clear. There may be a tendency for laboratories using the bioassay to attempt to infer mechanisms of action from the data. In some cases, this may be appropriate (e.g., with estrogenic compounds). With other responses or response patterns direct mechanism may not be as clear. Therefore, it is important that the limitations of the bioassay, and its purpose as a Tier 1-level bioassay, be emphasized. In combination with other tests (e.g., *in vitro* receptor studies; gonadal steroid production *in vitro* assays), there is strong potential for the data from the fish short-term reproductive to increase in interpretive power. There is no doubt that the methodology can be adapted by other laboratories; in fact, there are examples in the literature of laboratories not associated with the EPA or the EDSP that have already used this specific or adapted fathead minnow adult reproduction test (e.g., Kovacs *et al*. 2005. *J. Toxicol. Environ. Health A*. 68:1621-1641; Rickwood *et al*. 2006 as cited). Challenges include health of the fish and ensuring pre-selection of required male/female ratios for the test. Problems with fish health and improper ratios can result in some difficulties in data interpretation. Overall, however, the test appears to be robust in relation to these concerns. Appropriate guidance and options have been provided for the various endpoints. In some cases, significant training and interlaboratory sharing of knowledge and skills is required (e.g., histological assessment, including preparation of gonad tissue and assessment of reproductive status) to ensure that the data are as standardized as possible, even when it is qualitative in nature. In others, such as with the enzyme-linked immunosorbent assay (ELISA) and radioimmunoassay (RIA) protocols, it is clear that good intralaboratory QA/QC (quality assurance/quality control) does allow the detection of treatment differences even when absolute values are not consistent in interlaboratory comparisons (McMaster *et al*. 2001	Agreed. Tier I assay results will be considered in light of the entire battery.

as cited). However, the similarity of responses in fathead minnow (as well as other fish species) using much-studied compounds (such as estradiol/ethinyl estradiol) reinforce the validity of use of biochemical endpoints such as plasma steroid and plasma vitellogenin. Suitable guidance has been provided on the best available methods to use, e.g., GC-MS (gas chromatography-mass spectrometry) to measure water concentrations of exposure compounds rather than RIA/ELISA; plasma vitellogenin in preference to vitellogenin expression levels, etc. It is likely that as methodologies develop and become more standardized that guidance on the best available methods may change; however, at this time, appropriate guidance has been provided. A good example in this regard is the improvement in vitellogenin data with the development of fathead minnow-specific vitellogenin antibodies for use in standardized ELISAs. Five points are raised in regard to the statistical analyses which require clarification: (1) It is not clear what the units of replication are for the data. Is n=4 (number of tanks/ treatment) the unit of replication? Has a nested ANOVA been used to allow the fish to be proper subgroups? Using fish as units of replication otherwise is pseudoreplication. (2) Have power analyses been done to determine the ability of the test to minimize the chances of a Type II error (i.e., that no difference is concluded when in fact there is one; generally the result of high variability and/or low sample sizes)? This would seem to be especially important if an effect/no effect conclusion is the basis on proceeding to a second tier of screening. (3) Are the cumulative egg data analyzed by an ANOVA on the final day? This may miss some statistically significant data, e.g., when the slope of the change has changed between control and treatment, indicating a change in reproductive pattern. A two-way ANOVA may be a better statistical representation of the data. (4) Gonadosomatic indices are covariates and should be analyzed by ANCOVA (analysis of co-variance). When the assumptions for ANCOVA are not met, the regression lines can be examined independently. The data can be reported as GSI for presentation purposes (table or graph). (5) Using alternate statistical approaches may allow enhanced ability to use the data analysis to represent statistical and biological effects, with the intent to avoid making inferences of believed biological effect even when no statistical difference is indicated. Isn't more research required to conclude that a non-significant, X % decrease in fecundity is biologically significant enough to be an indication of endocrine/reproductive dysfunction (and, therefore, that is appropriate for a chemical to proceed to a second tier of screening)? Adequate guidance is provided on reporting the results; however, data interpretation is challenging if "trends" rather than statistical significance are considered to be biologically significant. With more descriptive endpoints, including secondary sexual characteristics and histopathology, good criteria must be provided to ensure consistency among laboratories. The provision of primary and secondary diagnostic criteria for histological analysis is a proper aid to this end. Solvent control issues must be reported (i.e., if solvent controls behave in a manner different from water controls). The interpretation of the data needs to be made within the proper context. Fertility endpoints can indicate whether male reproductive status may have been altered and warrants further study. Changes in fertility indicate whether there are viability issues with egg, sperm or both. However, the endpoint may be redundant in a Tier 1 screen if good histology data is available. As mentioned previously, clarification of and focus on the statistical analyses would improve the guidelines and data interpretation. Continued refinement and standardization of the biochemical endpoints [e.g., development of high-level QA/QC methods for small-volume plasma RIAs and ELISAs will strengthen the robustness of the data. Ensuring effective delivery of test compounds through appropriate guidance on solvent delivery and/or solvent-free delivery is necessary to avoid problems with test chemical delivery negating exposure validity.	The tank is the unit of replication for the endpoints, although some measurements are taken on a tank basis (i.e., fecundity, normalized to # female reproductive days) whereas others are taken per fish (i.e., tubercle measures) and averaged per tank. Post-hoc power analyses were included as part of the interlaboratory study. Cumulative egg data were analyzed on the final day. Consideration will be given to alternative analysis. The use of ANCOVA for evaluating GSI will be considered. The intent of the statistical evaluation component is to avoid unfounded inferences of biological effect, although using purely statistical approaches risks throwing out biologically relevant results (e.g., males induced to produce VTG can look like outliers from a purely statistical perspective). Because this is a screen, reproductive measures that differ statistically from that of the controls are considered a flag to be properly weighted in a weight-of-evidence determination across the Tier 1 battery. Agreed. This is not specified in detail in the protocol and additional clarification will be added. Agreed, additional guidance will be added to the protocol.

The high dose often appears to be adequate to elicit a response. There are a few instances where doses other than the high one resulted in effects (e.g., female gonadal atresia in ketoconazole exposure; interlab study) or where high and low concentrations have opposite effects (e.g., ethinyl estradiol in peer-reviewed literature). It *may* be possible to limit future testing to high concentrations. Is a dose response a necessary criterion in this screening assay (as it is in other toxicological studies), and to what extent are two or more doses required?
Reproductive behaviours could be an area in which more research could be done "down the road". As more studies or tests are done using the optimized bioassay, data could be accumulated which could lead to better understanding of "normal" behaviour. This "normal" behaviour could be useful for checking that behaviour of the control animals is within the expected ranges as well as identify altered behaviours which are the result of chemical exposure.

MM/ I think that the protocol does describe the methodology clearly so that the laboratory can comprehend the objective. In the assay initiation pre-exposure methodology, they state that additional tanks are set up to account for the lack of spawning of some fish or mortality etc. Fish whose gender could not be determined were excluded from the assay. However in the interlaboratory studies with the set compounds, sex determination errors were made a number of times preventing the optimized sex ratio from occurring a number of times in at least 2 of the 3 laboratories in the interlaboratory studies. Is this common? If these are three of the best labs in the US, what is going to happen when a number of other laboratories set up to run these tests? Is this a major problem? I was really surprised by this. Can the age of the fish used be adjusted slightly to help prevent this from occurring? With a territorial fish such as the fathead minnow this could have been a significant effect on the reproductive responses of the fish to the chemical and could very possibly alter the decision or increase the variability making the test less sensitive.
The protocol also states that

> 90% survival in the controls and successful egg production in controls. Spawning occurs at least every 4 days in each control replicate, or approximately 15 eggs/female/day/ replicate. Fertility > 95%.

However, these were not always the levels stated in the interlaboratory studies. Either the protocol changed somewhat after the interlaboratory studies were conducted or the protocol was not followed properly or was not clear enough.
The protocol also states that pre-exposure observations will occur in the same system/tanks as will be utilized for the chemical test (e.g., fish will not be transferred between tanks between the pre-exposure and exposure periods, which could induce stress).
What is there a pre-exposure period for then if the tanks cannot be moved? Most systems have limited space for tank placement. If tanks are not moved, how are the most similar spawning groups selected for the study? Some laboratories did move tanks into position in the exposure set up following the pre-exposure period.
It appears in some of the interlaboratory tests that control egg production was significantly greater than dosed tanks prior to chemical addition. They were starting the exposure with an effect on egg production prior to test solution addition. I think this selection of the most similar egg producers prior to test addition and randomly assigning these tanks to the various concentrations is one of the most critical steps to this assay. I don't think that it is clear enough in the protocol for all laboratories to complete.
The protocol states that 'The exposure phase will be started with sexually dimorphic adult fish from a laboratory supply of reproductively mature animals. Based on the technical judgment of experienced laboratory personnel, fish will be reproductively mature (namely, with clear secondary sexual characteristics visible) and capable of actively spawning. Apparently the experienced laboratory personnel could not select appropriate mature fish as sex ratios were

The need for more than one test concentration was questioned by one of the reviewers (DM). While this might be possible for relatively well-understood chemicals (which is most of what has been tested so far), to use this approach would be inadvisable for chemicals for which little is known about basic toxicity. Since test concentrations are based on short-term "range-finders", it would be relatively easy to "lose" a test if the longer-term 21-d exposure were conducted at an inadvertently lethal concentration. Also, even if mortality does not occur, it is quite possible that testing an animal at concentrations approaching lethality would cause enough stress to mute any type of endocrine-mediated response, which would be apparent at a lower concentration. Finally, there has been concern expressed in the scientific community for unusual dose-response curves (e.g., "U"-shaped) for endocrine-active chemicals which make multiple test concentrations/doses prudent, even for screening assays.

There were several comments concerning mis-sexed fish. This is a problem that generally becomes minimal as (a) laboratories become familiar with sexing adult fathead minnows, and (b) culture conditions are optimized to produce consistently mature fish. For example, in a recent experiment with more than 400 fish at the Duluth EPA laboratory, only one (0.25%) was mis-sexed. In any case, when using a group-spawning design in which one of the criteria for starting the test is that the tanks are spawning at a pre-determined acceptable level; the issue of mis-sexed fish in terms of affecting test outcomes should be minimal. Since what usually happens when problems arise is inadvertently calling immature males females, and there are multiple females in each tank (four), as long as egg number is normalized (at test end) to number of females, the test results should still be valid, even if only three of the fish were females.

often wrong. This can be a big problem as well. Can the age of the fish be adjusted somewhat to reduce the chances of this occurring? I thought if sufficient numbers of fish were in the pre-exposure phase that this would not be a problem. Should the number of pre-exposure fish increase? I feel that this is also a critical part of the assay as with the territorial fathead minnow, differences in the numbers of males and females can be critical. In fact that assay development looked at the influence of this and found that it did make a difference and that 4 and 2 would optimum. 'A randomized complete block design (4 blocks with one replicate of each treatment) will be used for the reproductive assay. This design is intended to randomize out the effects associated with the local environment (i.e., light and water) and possible trends associated with the diluter during testing. All fish will be impartially assigned to tanks before pre-exposure, then tanks will be randomly assigned to treatments within a block after spawning is established in the pre-exposure period. The blocks are filled in a random order, with the four tanks with the highest per-female fecundity (established during pre-exposure) being assigned first, followed by the second-highest spawners, etc. Thus, when one evaluates the difference between treatment means, the variability associated with experimental environment, experimental containers, and organisms being treated is removed and only the effect of the treatment remains'. This is pretty clear to me, but it is quite clear from some of the interlaboratory data that this is not that clear to some of the laboratories or not that easy to follow. Some of the studies had clear differences in egg production between control and exposed fish prior to chemical addition. Therefore, not only the effect of the treatment was being tested. Is this going to be a problem when a number of laboratories are set up to deal with the large number of chemicals that need to be screened? Additional exposure chambers should be set up for pre-exposure to account for a lack of spawning in some chambers and/or mortality during the pre-exposure phase. Any specimens whose sex cannot be identified will be excluded from the assay. For each assay, successful pre-exposure (suitability for testing) is established when regular spawning occurs in each replicate test chamber at least two times in the immediately preceding 7 days and egg production exceeds 15 eggs/female/day/replicate group. This is not what the table above states. There are different requirements stated in the different documents in this review. May be confusing to the laboratories as well, in fact the interlaboratory labs use different criteria. Procloraz Lab Wildlife International. During this phase, suitability for testing was established when regular spawning occurred in each replicate test chamber at least once in the seven days immediately prior to test initiation. The top 16 performing spawning groups were selected for the chemical exposure. It says clearly above that regular spawning must occur at least 2 times not 1 as stated in Lab B protocol. Also nothing is stated about 15 eggs per female. Top 16 selected. Were they the 16 closest or just the top 16? Some tanks may have been really great spawners and potential outliers. Should these be included? I do not think so. Would it not make more sense to select the 16 spawners that are closest to the average number of eggs for the whole pre-expsoure group? I think this may take out some of the variability found in some of the interlaboratory studies. I also think that some labs did not assign the tanks properly. It looks like for one study at least, all 4 of the best spawners were in the control tanks. Mean measured concentrations for that lab in the high concentration were lower than the medium concentration. I think that the numbers in their table must just not be right. Procloraz Lab B - The exposure system was operating properly for four days prior to study initiation to allow equilibration of the test substance in the diluter apparatus and exposure aquaria. This suggests to me that the fish were moved into aquaria for exposure period after pre exposure. This is not what is described in the optimized protocol. During this phase, suitability for testing was established when regular spawning occurred in each test chamber every 3 to 4 days. The top 16 performing spawning groups were selected for the chemical exposure.	The pre-exposure period addresses a number of needs, including establishing health and appropriate fecundity of the fish. Groups should remain the same to allow acclimation within groups, but tanks can be assigned to various exposure levels to have equal distribution of best spawners at test initiation, as clarified in the protocol after the interlaboratory exercise.

Treatment levels were randomly assigned in the exposure system and spawning groups were assigned to treatment groups using a rank-order approach. It seems like they used different criteria. Fish were moved into tanks apparently random, but if you look at the preexposure data for egg production this is not true. Exposure concentrations for this particular study ranged from 370 – 88 for the high concentration that was supposed to be 300, 130 - 32 for the medium 100 concentration and 26 - 5 for 20 lowest concentration. This is not acceptable and could explain some of the different or not consistent responses between laboratories. Is this just a poor system design? Are other laboratories going to have this kind of problem? Should there be a standard exposure system (peristaltic pumps vs diluter systems)? It appears that laboratory B and C had a great difficulty getting fish of similar egg production spread out among the various treatments. Quite often they had significant differences in egg production even before exposure was initiated. Laboratory C also added toxicant to exposure tanks for days prior to test start so fish must have been moved the day of exposure start. Methods state that you should not. Although I think if you are going to select spawners of equal potential this is required. Even if it is just moving the tank in the exposure system.	Protocol instructs "…using a proportional diluter or other appropriate delivery systems." Performance criteria for maintaining consistent exposures and acceptable water quality conditions will be imposed.
Are there any statistical tests or descriptions to look at differences in behaviour of the fish during the test? Altered aggression of males? Feeding abstinence etc. These changes can be critical in the evaluation of other endpoints such as decreases in weight, growth etc. The methods state that these are recorded but nowhere are they described after the study is finished. What is the appropriate magnification for examination of the eggs for fertility? Typical GSI for females are 8-13%. With that much variability in control fish, is this an endpoint that you expect to show effects with exposure? Are fixed ovaries weighed or are they weighed at time of sampling then fixed? If gonads are fixed in the body cavity prior to weighing the tissue, how can you be sure that the gonads are equally fixed and will not influence the weight of the gonads between fish? Why not dissect out tissue and then weigh prior to fixation? I do not think that stage of development and gonadal histology will be altered that quickly. These methods appear to be different between the documents in this review package.	Agree. The protocol will be clarified so fixation should be first, as removal of gonads prior to fixation and weighing could introduce artifacts that would impact histopathological evaluation and interpretation.
Histology – frequencies of the various cell types are recorded only? Are sizes deterimined? Presence of intersex? How many screens of cells are counted? Or all of the cells in the six sections are counted? Steroids – descriptions of why fish were used was because different androgens are important such as 11-ketotestosterone. Why is it not measured in these fish studies? Not sufficient blood? Generally, plasma samples are extracted prior to analysis. Is this done? When extracted, generally they are resuspended in 1 ml of buffer. If it is done this way, generally there is enough to assay for two steroids in duplicate. Where are these methods and why is 11kt not measured in males? Why is blood sample handling and treatment different between the two suggested bleeding techniques? One technique spins for 3 min at 15000g with aprotini and the other spins 5 min at 7000g without aprotinin and done at room temperature. This should likely be similar between sample collection methods I would think. In the text for gonad removal, the gonads are fixed in the body cavity. In the described methods in appendix f, the gonads are removed without being fixed. These should be the same and which one is better? I prefer the fixing after removal and weighing to eliminate potential weighing issues. Sampling – it is stated in appendix f that sampling should start with reference and increase in toxicant concentration. Is this normal? Should one not randomize sampling as well? Should it be done blind? Can sampling this way bias your subjective measurements?	In response to the question on 11-ketotestostrone (KT) in test animals, there are two reasons why this wasn't viewed as practical/necessary. First, as opposed to estradiol or testosterone (T), there is a lack of commercial availability for some of the reagents needed for a KT RIA (tritiated KT, antibodies), so it is not an easy measurement to recommend for routine use. More importantly, based on a relatively large number of control studies and experiments with test chemicals conducted at the Duluth EPA laboratory, KT and T concentrations in male fathead minnows appear to be consistently, positively correlated, suggesting that T status is a reliable indicator of KT for most situations (e.g., Jensen et al. 2001).
It is clear from the results of the interlaboratory studies discussed above that either the protocol is not clear enough in all aspects for the laboratories to conduct without problems or that the laboratories themselves were just having problems. It is also clear from some of the references to discrepancies in the techniques recommended in	This is referring to analytical sampling and is standard practice to avoid residue of previous samples contaminating subsequent samples, which would have a much greater impact if a

different appendices and different documents that this should also be cleared up to help laboratories to carry out this assay as consistently as possible. The individual appendices on each of the endpoints such as male secondary sex characteristics are critical to the correct observation and measurement of a number of the endpoints to ensure consistency in the interlaboratory studies. This is difficult as a number of the endpoints have a subjective factor in the measurement and these detailed documents are required. The histology document is also very detailed but the methods state that an experienced histolopathologist is required. Will experienced histologists actually follow the histology appendix? I think carrying out the exposures is the hardest and most critical part of this assay. They have done a very good job at detailing the observations and measurements for this assay. For a regulatory test such as this it is critical to have consistent compilation and statistical procedures. The Canadian Environmental Effects Monitoring program has created a web based data input, data compilation and data statistical analysis site. Industry or their private consulting firm input the data for a wild adult fish survey and a benthic community survey. Once data is inputted, the program looks for outliers prior to conducting the various required statistical analysis. A similar type of consistency should be created for reporting on the Tier 1 battery of tests. The website that describes the electronic reporting for this program is: http://www.ec.gc.ca/eem/english/ppv3_software.cfm There are no real explanations as to what is done with the behavioral descriptions during the test. Changes in territorial behavior etc. How are these subjective endpoints evaluated statistically? I found it fairly difficult to actually get the data out of the interlaboratory lab final reports. If warranted, please also make suggestions or recommendations for test method improvement. These suggestions are littered among the above discussion. **HS/** The ISR itself provides more a summary description of the Fish Assay and the endpoint methods (p. 16 ff), without giving detailed information. Such detailed information is given in attachments F and G. Still, a few points remained confusing to me: - on p. 81 of the ISR, the following endpoints are indicated as "core endpoints of the Fish assay: fecundity, histopathology, GSI, sex steroid measurements, vitellogenin and secondary sex characteristics. Surprisingly, the GSI is not included in the list of test endpoints on p. 22 ff (but in attachments F and G). On the other hand, the list of endpoints on p. 22 ff as well as in attachments F and G indicate further endpoints, for instance fertilization success („ fertility"), or behaviour which, however, were not measured in the validation tests. Why was the potential of these endpoints not further evaluated, be it by an in-depth literature discussion or by practical measurements ? - The endpoint fecundity seems to suffer from extensive variability, as indicated repeatedly in the IRS. Thorpe et al. (2007, AT 81 :90) in which they claim that they could substantially reduce variability by using a modified egg collection procedure. Why was no reference given to this approach – is it not effective or is it applicable to the group breeding protocol as used in the Fish assay ? - concerning the endpoint VTG, I would consider a more detailed discussion on the potential pitfalls of the ELISA methods – which are most widely used for VTG analysis – to be helpful (see also below). The validation of an ELISA for a specific application can be critical (see, for instance, the discussion between Myhlchreest et al., 2003, CBP 134C:251, and Tyler et al., 2004, CBP 138C:531). The ISR simply states that the availability of commercial kits is likely to improve reproducibility. I agree on that, but even with commercial kits there will be a number of potential pitfalls. It would be very helpul if the experience available in the EPA laboratory on the use of VTG ELISAs would	high-concentration sample were measured before a low or control sample. Correct. GSI was left out of this endpoint description in the ISR. This omission was an oversight; GSI is an important endpoint and should be evaluated in accordance with the protocol. Fertility was included in the interlaboratory trials. Behavior parameters have not been established for measurement, as this is primarily intended for gross evaluation of abnormalities in behavior if they are observed as general signs of toxicity.

flow into either the ISR or the assay protocol in order to prevent newly starting labs from repeating failures and mistakes. While technical descriptions in the ISR and attachment F are not sufficiently detailed to be able to conduct the assay, to measure the prescribed endpoints and to statistically analyse the data, attachment G provides a detailed description on blood sampling, vitellogenin analysis, sex steroid analysis, tissue preparation for histology, fish maintenance and particularly for chemical dosing and statistical evaluation that should be satisfying for those purposes. Particularly helpful is the decision tree for selecting statistical tests, as provided in attachment G and I recommend that the ISR explicitly refer to this (currently, the statistical section in the ISR – chapter 4.6. – goes not much beyond some general textbook statements). Additionally helpful would be a pictorial guide to the secondary sex characteristics and their possible changes (it is partly contained in appendix C of the study plan in attachment E). Concerning histopathology, attachment H provides pictures on normal and altered gonad morphology, what should be extremely helpful for laboratories performing the assay. As far as I know, also OECD is working on a histopathology atlas for FHM, zebrafish and medaka, what would further support the reproducible application of the endpoint "gonad histopathology".	
4) The intra- and inter-laboratory reproducibility of the test method should be demonstrated. **BEB/** Yes and no! Generally speaking, the outcome of the tests was very satisfactory. Somewhat surprisingly there were, however, some inter-laboratory differences in how they carried out their tests, which may at least partly explain somewhat confusing results, in particular with more "difficult" substances (e.g. SDS). **RDG/** Overall, results obtained appear sufficiently repeatable and reproducible for the purposes of this screening assay. The results from the optimization studies are generally reasonably consistent, particularly considering that the various experiments for a given chemical were oftentimes conducted with different experimental designs, different exposure times and concentrations, and different endpoints. The validation studies provide a better basis for probing this question. For these studies, three laboratories conducted studies with five chemicals - 4-*tert*-octylphenol (4OP), prochloraz, ketoconazole, vinclozolin and sodium dodecyl sulfate (SDS). For each chemical, the same three exposure concentrations were targeted by each laboratory, and the same endpoints were measured (fecundity, gonadal histology, gonadal-somatic index – GSI, estradiol – E2, and VTG in both sexes; in males, also number/appearance of tubercules and size of dorsal fat pad). Since for theses studies, identical designs were employed, differences among experiments reflect inherent biological variability, inherent assay variability, and/or inherent or human error-associated laboratory variability. Only the last possibility would seem readily fixable. Overall, the collective results for a given compound (except SDS, probably a poor choice for this purpose, as explained, p. 79), were sufficiently consistent across the three laboratories. However, considerable, and at times surprising variability, was observed for a given endpoint. For example, in females exposed to 4NP, gonadal atresia was seen at all three exposure concentrations by Lab A, only at the highest concentration by Lab B, and in no exposures by Lab C (p. 69, Table 5-1a). VTG was reported to be significantly elevated at the medium and high exposures by Lab A, while no effect was reported by Labs B and C. Similar endpoint variability is noted for females exposed to ketocanozole (p. 74, Table 5-3A) and for males exposed to	Overall, the reproducibility was deemed by the reviewers to be acceptable. Some recommendations made to improve the protocol should help reduce variability in some of the endpoints even further. In summary, this criterion is judged to be met.

vinclozolin (p. 78, Table 5-4b). However, it does not appear that there were major inconsistencies when the results are viewed collectively. That is, there does not appear to be a case where for a given chemical, one laboratory was at complete odds with the others concerning the potential for the chemical to perturb reproduction via the HPG axis; all are suggested as having that potential. However, all four of these chemicals are known HPG axis disruptors, so it would be disturbing if at least the highest exposure concentration had no significant effect(s). The other relates to an issue discussed earlier under question 2 - that it is not clear how results from this assay actually will be used to trigger more involved testing (Tier 2). Clarification of this would help inform an analysis of the importance of the inter-laboratory variability demonstrated for various endpoints in these validation studies. **DM/** The interlaboratory comparison, in addition to similar tests published in the peer-reviewed literature, provides adequate data to confirm the repeatability and reproducibility of the bioassay and chemical test methods. An obvious example is the bioassays done with estrogens/weak estrogens which may represent the largest database (predominant responses include, e.g., decreased fecundity at high concentrations, decreased gonad size, and increased vitellogenin). It should be noted, however, that it *often takes more than one test to identify the patterns.* Whether this is a by-product of interlaboratory variability or the range of potential fish responses (e.g., based on fish genetics, prior life experiences, etc.) is not apparent at this time and requires further study. It is the *weight of evidence of multiple tests*, as well as the *potential for response patterns within tests due to effects on multiple endpoints related to the HPG axis* (e.g., changes in fecundity, gonad histology, and biochemical parameters) that provide the overall rigour of the test. **MM/** I was disappointed somewhat by the differences or lack of consistencies between the three laboratories. Very rarely did all three labs demonstrate the same responses in any of the endpoints. I realize that the studies are not designed to clearly indicate mechanisms etc, but I was fairly disappointed in the lack of consistencies. Three laboratories are also fairly small as well in terms of an interlaboratory study. It did demonstrate that some of the endpoints such as fecundity are quite robust in terms of fecundity etc., but I feel that with increased control of fish selection, number of fish available for the test etc. that the mis-sexed fish should occur less often. **HS/** From the data presented, I have no access to within test variability of the various measurements and endpoints, therefore the following comments focus on intra- and inter-laboratory variability. Given the inherent variability of reproduction-related parameters in fish, and the fact that some of the assay endpoints are still in an early stage of development/application (see above), the results obtained with this assay show sufficient repeatability and reproducibility of the Fish Assay and its endpoints. The results of the inter-laboratory study (Batelle 2005) nicely illustrate that, as always, reproducibility is better with higher doses and with clearly defined modes of action. Further improvements of repeatability and reproducibility may be achieved by further standardizing assay parameters. For instance, I wonder how similar or different FHM strains from different laboratories are. From zebrafish I know that differences can be substantial, but for FHM I have not seen much studies addressing that problem; also the ISR does not provide information on this issue. Further options to improve standardization could include feeding conditions of the fish. Assay protocols say that the fish are fed twice a day with brine shrimp, however, no specification on the amount or on the strain of *Artemia* is given, although *Artemia* strains can differ greatly with respect to nutritional quality and toxic burdens. Since the nutritional status of the test fish can strongly influence the HPG axis, intensive consideration should be given to the aspect of test fish nutrition. Altogether, however, these are minor comments. As said above, I feel that considering the inherent variability of fish reproduction, and	

comparing to other biological assays using fish, the FHM Fish Assay shows sufficient repeatability and reproducibility.	
5) Demonstration of the test method's performance should be based on the testing of reference chemicals representative of the types of substances for which the test method will be used. **BEB/** (a) Choice of test substances was generally good. In one case, however, the substance (SDS) turned out to be problematic due to instability in water and rendered further problems in result interpretation due to individual variation in test performance at the three participating laboratories. This is saying less about the shortcoming of the test but more about the importance that test protocols are to be strictly followed at all times. (b) As mentioned above, the chemical analytical methods resulted in inter-laboratory variations that seem problematic in a direct comparison between labs. It is quite clear, however, that individual lab results render similar levels of sensitivity even if absolute levels of e.g. VTG were found. The precision of these methods may be expected to improve by time, experience and method development. (c) As mentioned above under #3, it is preferable that the number of statistical methods is narrowed to a few and that those will be declared as mandatory, i.e. the same statistical test package should be applied by everybody. Also the borderline between the use of non-parametric vs. parametric test should be clear-cut in the SOP. **RDG/** Test compounds used in formal optimization studies included a strong estrogen receptor agonist (estradiol), several weak estrogen receptor agonists (methoxyclor, 4-nonlyphenol, 4-*tert*-phenlyphenol and bisphenol A), an estrogen receptor antagonist (tamoxifen), two androgen receptor agonists (methyltestosterone and 17□-trenbolones), several androgen receptor antagonists (flutamide, vinclozolin and p,p' DDE), a modulator of steroid metabolism (fadrozole), several "multi-modal" chemicals (ketocanozole, prochloraz, cadmium chloride, and fenarimol), and several chemicals with unknown mechanisms of action (atrazine, prometon, perfluorooctane sulfonate, and 3-benzlidene camphor), two complex mixtures (bleached kraft mill effluent and metal mining effluent) and two negative controls (i.e., toxicants considered to have no involvement with the HPG axis) – potassium permanganate and ammonium perchlorate. Collectively, these chemicals and mixtures provide a very substantive and appropriate basis for the design and development of this assay. The analytical and statistical methods employed were appropriate to demonstrate the performance of the assay in these optimization studies. Concerns for consistency in analytical methods recommended for future implementation are discussed above. **DM/** ***(a) Test substances*** Overall, the test substances chosen were appropriate (see question 2 above). An interlaboratory comparison of an estrogen antagonist (e.g., ZM 189,154) would have determined the ability of the assay to detect anti-estrogens. It seems a large gap that an anti-estrogen was not tested (and that there is limited peer-reviewed data using a comparable fathead minnow reproduction bioassay).	Overall, the reviewers agree that the chemicals used in the validation studies for the assay were appropriate. In conclusion, this criterion is judged to be met.

The interlaboratory comparison using sodium dodecyl sulfate (SDS) for the non-endocrine active toxicant was not a good selection given the challenges related to exposure. There is an indication that general whole-animal endpoints and fecundity were affected (Lab B primarily). However, the overall results do not support the conclusion that toxicity through non-endocrine-mediated mechanisms can be distinguished by the test. ***(b) analytical methods, and*** The analytical methods are appropriate. As mentioned previously, continued development of the methods to ensure good QA/QC should be an on-going goal, through, e.g., further refinement of RIAs and ELISAs, training and guidance documents in histological methodologies, etc. **MM/** For the interlaboratory comparison, I was surprised that the laboratories only "generally" followed the protocol. Should this not be a requirement? Especially for the interlaboratory comparison? Are three labs sufficient to demonstrate the potential interlaboratory variability?? Delivery of test solution - Even though these techniques were well described and were evaluated before the initiation of an exposure, the three laboratories experienced some difficulties in producing consistent stock solution concentrations and maintaining exposure levels with some of the test substances. Will this be common with these test systems? Surprising given that it was looked at closely prior to the studies but still it was not done well at all. See reference above for one lab and procloraz. Procloraz – laboratory B had some large variations in compound concentrations that should not really occur with an experienced laboratory. Surprising I guess. Laboratory A mis-sexed fish in all its treatment groups. This is not acceptable and should be corrected for. Suggestions should be made to prevent this in other laboratories. This also resulted in reduced control fecundity that would result in potential misinterpretation in other studies. They were lucky that the compound was active enough to overcome this increased variability. Although the interlaboratory study was truly interlaboratory for the exposure itself, the histopath was all done at one laboratory. A statement such as this "Significant treatment-related testicular degeneration was observed at the highest treatment level from Laboratories A and B, with possible treatment-related effects also observed from Laboratory C" would not be made if different histo labs had conducted the evaluations. Octylphenol – Lab A and C mis-sexed fish similar to the first procloraz exposure. Fecundity variability in controls from two of the labs were lower than normal. Again, should increased numbers of fish be used in the pre-exposure phase to select fish that are more similar. Vinclozolin – concentrations between the three labs is highly variable. Suggests that there are many places where errors can be generated. Low 75, 150, 84 this is too large of a spread. High 830, 1200, 760. Not good. I was disappointed with the negative control as the laboratories used only one concentration and that concentration was not the same at the three laboratories. ***Reliability*** is defined as the reproducibility of results from an assay within and between laboratories. Has the assay really been demonstrated as reproducible within a laboratory? This means the same substance was run a number of times within the same laboratory. Between laboratory reproducibility is really not that great. This is however, due to the numerous endpoints measured in the assay itself. With so many comparisons between control and exposed fish, it is very unlikely that any assay of any chemical will result in no significant differences. This is critical in evaluating chemicals and the battery of tests will have to line up with similar mechanisms being affected with the same chemical. For example if the steroidogenesis assay picks up effects on androgen production, then the short term fish assay should also demonstrate reductions in circulating androgen levels or at least corresponding reductions in expression of male characteristics.	

HS/ To evaluate the suitability of the Fish Assay, a series of positive substances with known mode of endocrine action, substances with multiple or uncertain modes of action and suspected negative substances were used. The selection of test substances to demonstrate (anti-) estrogenic or (anti-)androgenic activities was appropriate, also the selection of multimodal compounds since their results pointed to potential problems in interpreting the test results (see above, discussion on charge question 2). The results with the negative substances, however, are more equivocal. Permanganate was used at concentrations that exceeded lethal thresholds so that the results from this test are questionable (although the OECD Draft report/background reference 1 concludes from the permanganate study that the "21-day-fish screening assay appears to be rather robust and specific enough to discriminate between substances with endocrine active properties and other substances", p. 25). Perchlorate did neither affect reproductive nor biochemical endpoints but induced a significant increase in atretic ovarian follicles. Oocyte atresia is a rather unspecific response that can be caused by many stresses in addition to endocrine disruption. Thus, the results with perchlorate are rather ambiguous and the intention – to show the behaviour of a clearly negative substance in the Fish Assay – could be not achieved. Similar comments apply to SDS. Concerning the analytical methods for the various endpoints, state-of-the-art techniques have been used. One of the key endpoints in the Fish Assay is VTG. The available data indicate considerable variation for this parameter, both intra- as well as interlaboratory (see Batelle 2003b, Hutchinson et al. 2006). As pointed out in the IRS, this may be in part due to the use of different assays, antibodies etc in the indiviudal laboratories and it can be expected that with the availability of commercial FHM VTG kits, at least inter-laboratory variability will get smaller. There may be two further sources of variation in VTG determination: One source is blood sampling, which, as outlined in the ISR, is not easy with a small fish as FHM. A second factor not to be overlooked is that many ecotoxicological testing laboratories up to now rarely applied biochemical assays, so that it may need a certain training period until results will get more repeatable. The ISR does not give much hope that using real time RT-PCR instead of the ELISA would reduce variability of the VTG endpoint. Here, I have a different opinion. PCR avoids sources of variability that apply to the ELISA, mainly blood sampling (liver sampling is much more easy) and interference with other plasma proteins. Given the rapid development of PCR technologies during recent years, real time methods are getting highly repeatable – in my experience more repeatable than ELISAs. Thus, I recommend to re-consider the potential of qRT-PCR for measuring VTG in the Fish Assay. Another endpoint „under development" is histopathology. This is nicely illustrated from the "pre-" and "post-Heidelberg" results on gonad histopathology in attachment C. Also here, as with VTG, knowledge on fish histopathology develops quickly - see papers such as that of Leino et al. (2005) as well as the excellent histopathology guideline for FHM (attachment H) - so that repeatability should improve in the future, despite the currently low level of standardization. Having said so, I was surprised by the low variation of the histopathological results of the interlaboratory study presented in the ISR. However, as I understood, the low variability of these results is an artefact since histopathological examination for all three partners was done by one and the same laboratory (EPA). Thus, this study is more likely to show intra- than interlaboratory variation. Statistical testing is shortly addressed on p. 65 ff of the ISR. I am not expertised in statistics, but probably there will be not one fixed statistical approach on evaluating the assay data, but the approaches will have to be adapted to the endpoint and the data variability. Examples of "decision trees" on how to select the appropriate statistical tests are discussed in attachment G and in background reference 1. Thus, this aspect seems to be well covered. One question I have – but this may be related to my ignorance in statistics – is whether a n-number of 8 males and 16 females (in four replicates) per test concentration is not of a rather poor statistical power, particularly when considering the high inherent variability of reproductive endpoints. Or, to express it differently: do we loose	There were several comments concerning appropriate VTG measurement techniques. Our best guidance at this point would be to measure VTG protein levels in the plasma of the fish using an ELISA technique with standards and antibodies specific to the fathead minnow. Previous work has suggested that unreliable results can occur when using ELISAs developed for other species to determine fathead minnow VTG (e.g., Mylchreest et al. 2003 and Korte et al. 2004). In addition, virtually all the studies evaluating the fathead minnow 21-d tests have used the ELISA approach and, while it is possible that measurements of VTG mRNA (e.g., via PCR) might yield comparable results/conclusions as measuring protein, this has not been clearly established, particularly in females, where comparatively little is known about the kinetics of VTG mRNA expression and protein production under either control or chemically-impacted scenarios.

sensitivity because the assay uses comparatively low n-numbers of fish ? Were there ever any model calculations made to better define the cut-off between practicability and statistical power in deciding on the group/replicate size ?	
6) The performance of the test method should have been evaluated in relation to relevant toxicity testing data.	N/A
7) Ideally all data supporting the validity of a test method should have been obtained in accordance with the principles of GLP.	The interlaboratory study was performed under strict GLP guidelines in all 3 participating laboratories.
8) All data supporting the assessment of the validity of the test should be available for expert review.	All data used to support validation can be found at www.epa.gov/scipoly/oscpendo
Strengths and limitations: **BEB/** Yes, that is very well covered in the documents. A lot of supporting data are presented in the OECD documents. **RDG/** The discussion of strengths and limitations of the assay overall and specific endpoints are very succinctly summarized (pp. 82-85). However, this information, particularly for limitations, is presented essentially as bullets, with no discussion. Further elaboration may be useful, and some distinctions made. For example, the limitations of histopathology (additional time and services of a pathologist) and sex steroids (radio-immuno assays may be challenging for some commercial laboratories) appear on equal footing; this is unlikely the case. For this analysis, it may be useful to make time and cost estimates for the endpoints selected. **DM/** The strengths and limitations have, for the most part, been adequately addressed. Although it is lauded that an attempt to minimize animals has been made, it is also important that adequate power and the most robust statistical tests are used. In the long run, ensuring good power will minimize animal use even if animal numbers per individual test may be higher. Recent information supports the hypothesis that fathead minnow are relatively more sensitive to endocrine disruptors than other small-bodied fish (e.g., Kidd *et al.* 2007. *Proc. Nat. Acad. Sci.* 104: 8897–8901). Therefore, its selection as the species for the fish short-term reproduction assay is strengthened. Although listed as "cost effective, and…reasonably rapid", it should be noted that other fish reproduction tests could be used which may be shorter and cheaper (e.g., zebrafish). However, the inability of zebrafish and medaka to be used for the suite of biochemical endpoints severely limits their usefulness. Because whole organisms have multiple mechanisms to maintain homeostasis (e.g., compensation in steroidogenesis, metabolic clearance, etc.), whole-animal exposures may not indicate a biological effect that is evident in *in vitro* tests. As well, differences in vertebrate group or fish species responses/sensitivities/modes of action may result in a lack of effects in fathead minnow compared to other species. Therefore, the fish reproduction test may not be able to indicate an endocrine-active compound during the screen. However, this is balanced by the multiple tests to be used in the Tier 1 screens. **MM/** What is the estimated cost of the assay for one chemical? Is this really cost effective? I agree completely however that it produces a	In general, strengths and limitations are well covered. Some additional discussion by the reviewers on strengths and limitations are well taken. It is concluded that the strengths and limitations of the assay are well understood.

relatively large amount of data over a reasonably short period of time. It uses an intact HPG axis and has the ability to detect (anti) estrogenic and (anti) androgenic responses. Do sufficient numbers of laboratories exist today to conduct the required number of tests? What would the learning curve be for additional laboratories to develop the expertise for conducting this assay well? Are sufficient laboratories available to conduct the analysis on the biochemical endpoints? Vtg – agree with strengths of this assay for the induction in male fish. Due to the reduced amounts of plasma available, has an interlaboratory study been conducted with this endpoint? Is it possible to create a large pool of plasma for such a study? There appears to be two different handling procedures for blood collected from the caudal vein and from the heart puncture. These should be similar? Use of aprotinin? Are there going to be requirements to use a specific antibody for the fathead? Fecundity - fecundity as an apical endpoint, when combined with gonadal histopathology, provides a good indicator of reproductive health of the fish as impaired fecundity is an adverse effect with regulatory importance whether it is due to endocrine-mediated activity or another mechanism of action. This is very important as fecundity or number of eggs spawned can be influenced by other things than just endocrine disruptors. If a female feels that her young will not survive in the environment that she is in, then she may decide not to spawn her eggs. I agree that this is important from a regulatory perspective, but is not endocrine mediated. This endpoint was altered sometimes with negative chemicals in the OECD studies. How important is this that there are supporting endocrine responses that support an endocrine mechanism for the reduction in fecundity? Gonadal histology – May also be possible to respond in a mechanism un-related to endocrine disruption. Resorption of eggs due to an environment unsuitable for spawning could result in a histological change due to exposure. Overall, it is concluded from Phase 2 studies that the 21-day Fish Screening Assay, including the vitellogenin and secondary sexual characteristics endpoints, is relatively specific for endocrine active substances. However, care should be taken when evaluating the reduction in male secondary sex characteristics and spawning status, in light of other information available (e.g. other signs of toxicity, response on other endpoints, etc.). Vitellogenin measurement is a relevant, reliable and relatively specific endpoint for the detection of endocrine activity of chemical substances. Secondary sex characteristics are also relevant, reliable and relatively specific, but may need to be restricted to induction in female fish, and not reduction in male fish, to avoid false positive outcomes. Secondary Sex Characteristics – definitely one of the better androgenic endpoints in the whole fish assay. Although the specific result of reductions in male sex characteristics in males or the expression of male secondary sex characteristics in females on overall reproductive success are hard to estimate or determine, it is definitely a very good indication of endocrine alterations. Sex steroids – agree 100 % that reductions in steroids co-occuring with for example reductions in fecundity help support an endocrine mediated response. I agree that supporting responses are very important with this assay. Fathead as a species – ability to get sufficient amount of blood from the fish is a positive factor, however whole body homogenates are also possible for vtg and blood steroid analysis in smaller species. Benefits for Canada and US as it is a indigenous species. Limitations Must be a very well established laboratory to be able to culture and care for a fathead minnow culture. I am not sure how many laboratories will be available to conduct such assays for this regulatory requirement. If there is an answer for this, I would like to know it. What is the estimated cost of conducting a test of this nature? This should include technician time for the culture and maintenance of the culture. How many chemicals in what time frame have to been evaluated with the Tier 1 battery of tests? From the interlaboratory comparisons, I was quite disappointed that at least two of the laboratories incorrectly sexed the fish prior to	

exposure. This led to some tanks having the wrong ratio of males to females. For a species like the fathead minnow that demonstrates strong territorial activity, it can definitely influence the natural expression of endocrine mediated responses such as secondary sex characteristics. This was not addressed very well in the report and I feel that there should be a better way to address this problem. Would starting with slightly older fish prevent this? Having more fish to choose from in the pre-exposure phase?? I am pretty sure that this is an important issue with this assay and that it has not been addressed properly. Vitellogenin – how important is it to validate the Vtg assay kits between laboratories? If the same assay is used within the same test and comparisons are made to the control fish, then I am not sure that validation is required between laboratories. Unless the specific levels of vtg are used to rank specific compounds between laboratories for ranking in terms of Tier 2 testing and how quickly it will happen or if it actually does happen. Fecundity – are suggested number of additional pre-exposure tanks sufficient to ensure similar reproductive output exists prior to the start of the exposure. I think it is somewhat critical to attempt to ensure that fish of equal reproductive potential are used. This decreases variability in control fecundity data and helps to determine chemical effects. Should it not be the 16 tanks with fecundity results around the mean fecundity for all of the tanks. It is quite often that one tank really out spawns all of the others and is a real outlier. These fish should not be used in the study similar to the lowest fecundity ones not being used. In the interlaboratory studies some of the labs did not meet the performance criteria for control fecundity. Why is this? Should they be changed or should the labs have done the studies over? Histopathology – What is the estimated cost of the histopathology for this study? How limited are the number of qualified histopathologists? Is there a desire on their part to actually provide this type of service? Steroids – a number of laboratories are now using the ELISA procedure for steroid analysis. Is this an accepted protocol for measurement of steroids for this test? It would reduce the use of radioactive substances which are potentially more harmful to the environment. I agree with the limitations identified for this assay. I do like the commitment of the EPA to re-examine data after a number of substances have been evaluated to determine how well these tests are predicting effects demonstrated in Tier II studies. Are some negative compounds also going to be tested with the Tier II assays to make this re-examination complete? I agree with the discussion of genomic studies and their potential in years to come. There is still a great deal of research required however before any of these tools can be used to predict whole organism responses or in regulatory decisions. The samples produced by these whole fish assays however, provide a great potential for the rapid development of these tools in the fathead minnow. There should be a directed research aspect to this program to take full advantage of the studies being conducted. Ie there should be a number of research labs connected directly to this program and samples such as liver that are not being utilized for any endocrine endpoints be used to determine of the expression of such genes. **HS/** Strengths and limitations of the optimized fish short-term reproduction are discussed particularly but not exclusively in chapter 6.3 of the ISR. The arguments put forward in chapter 6.3 are well founded, including the reasoning why an in vivo test has to be used (intact HPG axis !). One aspect that in my view would have deserved more consideration is the question why this assay represents an „optimized“ assay. Chapter 4 deals with the optimization of the assay, and to this end, the ISR discusses (i) how the assay responds to chemicals that are known to interfere with the HPG axis and how it responds to non-HPG-interfering chemicals as well as chemicals with muliple actions, (ii) and how different test configurations influence the test outcome. The discussion is based largely on information from the published literature (with the inherent problems of data comparability between	

non-related studies using various test protocol modifications), on data from an OECD interlaboratory study (OECD 2005) and on data from studies performed by Batelle Institute in the frame of the ESDP programme. A first observation is that the data from the literature are in some cases surprisingly fragmentary what makes a conclusive evaluation difficult. For instance, if it comes to the strong estrogen receptor agonist, EE2, there appears to exist just one adult FHM reproduction test using the group spawning protocol, but no study using the non-spawning or the pair-breeding protocol (table 4-1). A further observation is that the discussion on the various possible test confirgurations is restricted mainly on the comparison of the spawning versus the non-spawning protocol (chapter 4.2), but neglects other possible assay modifications such as - group breeding versus pair-breeding, - composition of the breeding groups in the group spawning format (why specifically 4 + 2 ?), - criteria for selecting the core endpoints (why, for instance, has fecundity been selected as apical endpoint, why not fertility ?), - optimal test duration. On p. 20 of the ISR, the pro's and con's of the group spawning versus the pair breeding configuration are discussed. The primary drawback to the group spawning protocol is a principal one, i.e. the inability to associate the various endpoints to an individual animal. There may be other drawbacks of the group protocol, for instance, in the OECD Report on Phase 1B (attachment C), territorial behaviour is mentioned as a possible confounder in FHM group spawning tests. The primary drawbacks to the pair breeding protocol appear to be mainly technical ones (increased costs, compromises with numbers of fish and tank replication). Since variability of reproductive parameters such as fecundity and fertility can be high, as repeatedly reported in the ISR-cited studies, the promise of the pair-breeding protocol to reduce variability (see also Thorpe et al. 2007, AT 81:90) deserves careful consideration. Unfortunately, a direct experimental comparison between the group spawning and the pair breeding protocol (as for the spawning versus non-spawning protocol) has apparently not been done in the ESDP programme, and also the literature provides no comparative data on the two test configurations. Such a comparison could have clarified a) whether the suspected conceptual advantages of the pair breeding protocol over the group spawning protocol come true, and b) whether the conceptual advantages outweigh the technical disadavantages of higher costs/lower tank replication. Another aspect that would have deserved more attention is the potential use of fertilization success („fertility") as core endpoint. The optimized protocol considers as apical endpoint only fecundity. This endpoint reflects primarily disturbances of the HPG axis in the female, perhaps indirectly also in the male (for instance, altered male mating behavior could modulate female fecundity). With fertility, a more direct access to assessing chemical impact on the HPG axis in the male fish would be possible. In my own experience with zebrafish, fertility was usually the more robust (less variable) endpoint than fecundity, and thus was also more sensitive in detecting EDC effects. Of course, FHM can be very different to zebrafish, still, data provided in attachment D seem to point to a lower variability of fertility compared to fecundity. For instance, in the experiments of the Springborn Smithers laboratories, egg number per female as well as spawns per female were the most variable endpoints (CV > 80 %, p. 33) whereas CV of fertility was < 10 %. A more in-depth evaluation of whether fertility would provide a more robust endpoint than fecundity appears to be worthwile. A screening assay should be fast. The optimized Fish Assay goes for a 21-day-duration. There have been discussions whether a 14- rather than a 21-day-duration may be sufficient for the purpose of the assay, but the ISR gives relatively little consideration to this discussion. There are some comments scattered over the ISR, but there is no focused discussion on the pro's and con's of 14 versus 21 days. From the published literature, it is difficult to come to a conclusions, as most studies differ by more parameters than test duration, what makes it difficult to conclude on the importance of exposure duration for the	One comment by HS was raised relative to group versus paired spawning designs, and is similar to a question raised by RDG earlier (on validation criteria 3 – test protocol). While fathead minnows certainly are amenable to paired-spawning, and this approach is definitely useful when trying to attribute spawning success to an individual animal, the group design does offer some advantages. For example, as noted above, this would help obviate possible problems associated with mis-sexed fish. Probably most importantly, use of a group design simplifies the test in several ways; specifically, it minimizes the number tanks needed to achieve a reasonable statistical power (i.e., four as opposed to 12 to 16 pairs) resulting in (a) fewer tanks to purchase, maintain and clean, (b) fewer chemical delivery lines (and pumps) to maintain, and (3) fewer chemical analyses required for exposure verification. Depending on the test chemical, this latter point alone could save several thousand dollars over the course of a test. So, to keep the test as simple and inexpensive as possible, a group versus a paired design seems reasonable. One reviewer (HS) suggested that the 21-d test could be shortened, for example, to 14 d to help optimize resource use, and make the test more consistent with the concept of a screening assay. While this would be desirable, the option has not been thoroughly evaluated from a technical perspective. Based on available data, while it is likely that exposures shorter than 21-d could be effectively used for very potent endocrine active chemicals like steroidal estrogens and androgens (e.g., ethinylestradiol, trenbolone, methyltestosterone), it is uncertain whether effects of weaker chemicals would be observed in 14 d. However, this is an important option to be evaluated with future comparative studies.

test outcome. For biochemical parameters such as VTG, a 14-day-exposure appears to be fully sufficient to result in an induction response (e.g., Panter et al., 2002, ETC 21:319). Ankley et al. (2001, ETC 20:1276) found that a 12-day-exposure period was sufficient to alter gonad histology, GSI, plasma androgen levels, etc. Support for the 21-day-duration seems to come mainly from assumptions on bioaccumulation kinetics (p. 22) and from the Batelle (2003a) study that found (i) that effects of 17beta-trenbolone did not differ between the 14- and 21-day-protocol, (ii) flutamide effects were consistent between the two exposure periods, except that effects on androgen levels were observed after 14 days in the high dose group only, while after 21 days, also the low dose group displayed altered androgen levels, and, similarly, (iii) that fadrozole altered plasma E2 and VTG after 14 days only in the high dose group, but after 21 days also in the low exposure group. Thus, the Batelle results indicate that the response patterns are basically similar betwen 14 and 21 days, but that the longer exposure period makes the assay more sensitive – what is clearly an argument for the 21-day-duration, since a screening assay should be sensitive (more sensitive than specific). It would have been advantageous to summarize the arguments for and against the 21-day-duration anywhere in the ISR, for instance, as an own sub-paragraph under chapter 4 (protocol optimization). An "optimization" aspect well supported by the discussion in the ISR is the decision for the spawning instead of the non-spawning protocol. On p. 39 as well as on p. 60 ff, reference is given to comparative studies showing that several endpoints are responsive to EDC exposure in the spawning protocol but not in the non-spawning protocol. These observations make sense as the negative impact of EDCs should be more easily visible in an activated HPG system than in non-active one. In conclusion, coming back to a statement made in the beginning of this section, the available database is rather small, and thus, each decision fo a specific assay configuration must be at least partly arbitrary. Given this difficult situation, and given the fact that we cannot wait for another 10 years of research on test optimization, the "optimized assay protocol" as suggested in the ISR appears to represent a reasonable and sound compromise to the stated purpose of the assay.	
Performance criteria: **BEB/** Yes! Everything suggested has an impressing back-up in documentation from OECD activities in particular where Dr. Ankley has been one of the most prominent and experienced contributors. I can not remember any similar test proposal which has the similar high level of scientific back-up as the present assay proposal. **RDG/** This may be an important weakness of the assay protocol (if I am interpreting this question correctly). Very little information is provided on variables such as expected value ranges and biological and assay variability for the various endpoints. Thus, little information concerning "performance criteria" is provided. Such information concerning expected values and variability would be helpful, especially for laboratories initiating studies of this nature. A table with these data for male and female control fathead minnows and minnows exposed to model compounds (e.g., data from the studies used to optimize and validate this screening assay) would be useful for quality control, and for data interpretation in future studies. For example, it will be important for a laboratory to know in the case of highly variable data (in which the variability excludes statistical significance despite a large difference in central tendency values), if the variability observed is expected or suggestive of analytical error. Another issue that appears to merit clarification is that of statistical significance, which seems to be the key determinant of whether or not a result triggers further testing. Is $p < 0.05$ to be used? Or will the p-value vary, depending upon known variability with an endpoint? With respect to histopathology, when is a particular response (e.g., ovarian atresia, testicular degeneration, increased spermatogonia) deemed	It is agreed that performance criteria need to be better defined. More detail will be provided on performance criteria in revisions to the final protocol.

significant? **DM/** The animal husbandry requirements for the assay are well defined and represent good conditions for quantifiable reproduction in this species in captivity. The requirement for extra tanks during the pre-exposure period to ensure adequate sample sizes for the exposure and to limit the range of variability is important. Good attention has been paid to the challenges related to ensuring chemical exposure close to nominal levels as well as the selection of appropriate test concentrations. Compounds with low water solubility requiring an alternate method of dosing (e.g., oral administration) requires additional guidance. Monitoring survival, including in the controls to determine if there are non-test-chemical reasons for mortality, is important to avoid testing at acutely toxic levels. For additional endpoints, including behaviour, fecundity, secondary sexual characteristics, gonadal histology, and biochemical endpoints the appropriate parameters were chosen and guidance provided as applicable. The accumulated data from fathead reproduction tests (reviewed here as well as additional studies) provide a strong foundation of data to determine if subsequent tests are performing within acceptable standards. **MM/** Labs did not meet performance criteria for egg production in some cases. Are these too strict or should the study not be accepted?? With respect to the parameters selected, I feel that the most appropriate endpoints were selected and that the three laboratories were able to determine that the chemicals selected acted with similar modes of activity. This is very promising for the assay. If the different labs saw effects that were not of similar modes of action, ie some acting like estrogens and others demonstrating androgenic responses than we would be quite worried. Spawns per female per day demonstrate huge variability as an endpoint. Is it really useful to calculate? **HS/** Surprisingly, the question of performance/quality parameters of the Fish Assay is not explicitely addressed in the ISR. Of course, there exist performance criteria and methods of quality assurance for the Fish Assay, as they are laid down in some of the attachments (for instance. attachment G). However, it would be helpful to compile a list of the existing performance and quality criteria, in order to reveal where the Fish Assay is still lacking such criteria, as well as to stimulate a discussion how "good" these criteria are. Quality criteria exist at least for the fish test per se (for instance, in attachment G it is clearly defined what fecundity has to occur in a spawning group during the pre-acclimation period in order to be allowed to use this group for testing). For "newer" endpoints such as histopathology, they are partly still under development, for instance, the histopathology guideline says that control females should display only "a few atretic oocytes" without specifiyng what " a few" precisely means in quantitative terms. A serious problem to be discussed is the health status of the test fish. In the inter-laboratory study presented in the ISR, fish from at least one laboratory was infected by parasites and this admittedly influenced reproductive performance. Can diseased fish still be used for a valid Fish Assay ? Parasites can strongly influence the reproductive performance of fish, and, therefore, in my view, results form parasitized fish have to be discarded.	

Major Action Items

The protocol guidance will need some minor revision based on recommendations from the peer review panel. The principal recommendations include:

- Recommend that fish are as similar as possible in egg production at the beginning of exposure.

- Recommend that fish are sexually mature and of similar and optimal age for reproduction and to avoid mistaking immature males for females.
- Recommend clarifying guidance for equal distribution of spawning groups among treatments to avoid bias.
- Suggest clarifying use of behavior observations.
- Recommend additional guidance on methods for chemical delivery.
- Recommend a standard statistical approach (e.g., data analysis program or macro).
- Recommend additional guidance for data interpretation.

EPA accepts the recommendations and will revise the protocol guidance accordingly.

References

Ankley, G.T., Jensen, K.M., Durhan, E.J., Makynen, E.A., Butterworth, B.C., Kahl, M.D., Villeneuve, D.L., Linnum, A., Gray, L.E., Cardon, M., and Wilson, V.S. (2005). Effects of two fungicides with multiple modes of action on reproductive endocrine function in the fathead minnow (*Pimephales promelas*). *Toxicological Sciences* 86(2): 300-308.

Ankley, G.T., Jenson, K.M., Kahl, M.D., Makynen, E.A., Blake, L.S., Greene, K.J., Johnson, R.D., and Villeneuve, D.L. (2007). Ketoconazole in the fathead minnow (*Pimephales promelas*): Reproductive Toxicity and Biological Compensation. *Environmental Toxicology and Chemistry* 26(6):1214-1223.

EPA. 2002. A Short-term Test Method for Assessing the Reproductive Toxicity of Endocrine-Disrupting Chemicals Using the Fathead Minnow (*Pimephales promelas*). EPA/600/R-01/067-154.

Jensen, K.M., Korte, J.J., Kahl, M.D., Pasha, M.S., and Ankley, G.T. (2001). Aspects of basic reproductive biology and endocrinology in the fathead minnow (*Pimephales promelas*). *Comparative Biochemistry and Physiology C-Toxicology and Pharmacology* 128(1): 127-141.

Jensen, K.M. and G.T. Ankley. (2006). Evaluation of a commercial kit for measuring vitellogenin in the fathead minnow (*Pimephales promelas*). Ecotoxicol. Environ. Safety 64, 101-105.

Korte, J.J., E. Mylchreest, G.T. Ankley (2004) Comparative evaluation of ELISAs for detecting vitellogenin in the fathead minnow (*Pimephales promelas*)—a response to Tyler et al. Comparative Biochemistry and Physiology, Part C, 138(4):533-536.

Mylchreest, E., Snajdr, S., Korte, J.J., and Ankley, G.T. (2003). Comparison of ELISAs for detecting vitellogenin in the fathead minnow (*Pimephales promelas*). *Comparative Biochemistry and Physiology C-Toxicology and Pharmacology* 134(2): 251-257.