Fish Short-Term Reproduction Assay
OCSPP Guideline 890.1350

Standard Evaluation Procedure (SEP)

ENDOCRINE DISRUPTOR SCREENING PROGRAM
U.S. Environmental Protection Agency
Washington, DC 20460

6 October 2011

Table of Contents

Appendices

I. Introduction

This document was developed by EPA to provide guidance to EPA staff who will be reviewing the data submitted in response to Tier 1 Orders issued under the Endocrine Disruptor Screening Program (EDSP). This document provides general guidance and is not binding on either EPA or any outside parties. The use of language such as "will," "is," "may," "can" or "should" in this document does not connote any requirement for either EPA or any outside parties. As such, EPA may depart from the guidance where circumstances warrant and without prior notice. The SEPs are intended to be used in conjunction with the EDSP Test Guideline Series 890 and the Corrections and Clarifications document available on the EDSP web page.

This Standard Evaluation Procedure (SEP) provides guidance on how EPA generally intends to review studies conducted using the OCSPP Guideline 890.1350 Fish Short-Term Reproduction Assay (FSTRA) that are submitted to support requirements imposed under the U.S. Environmental Protection Agency's Endocrine Disruptor Screening Program (EDSP). The objective of EDSP Tier 1 assays is to characterize the potential of a chemical to interact with the endocrine system.

The product of the review will be a Data Evaluation Record (DER) that reflects how well the study conforms to the Guideline, evaluates how well the study and analyses were performed, and provides the conclusions supported by the data. The DER will include, for example, a list of any significant deviations from the guideline and their potential impacts, a list of significant information missing from the study report, a description of how the statistical analyses were performed and whether they were performed according to the guideline, and any other information about the performance of the study that affects interpretation of the data within the context of the EDSP. The DER should record details on all endpoints required by the guideline. The DER is intended to contain enough information to provide EPA with the ability to determine whether the study was performed according to the guideline and whether it is scientifically adequate.

The guideline recommends the critical materials, methods, and analyses that lead to successful performance of the assay. If a particular material, method, or analysis is specified in the guideline, it is usually because other materials, methods, or analyses are either known to be inappropriate, or at least have not been validated and there is concern for their potential influence on results. The Agency has posted Corrections and Clarifications on Technical Aspects of the EDSP Tier 1 Assays (OCSPP Test Guideline Series 890) in the docket; the link to this document may be found by way of the EDSP web page (http://www.epa.gov/endo/). It is therefore important to note deviations from specific materials, methods, or analyses in the DER, and provide the Agency's opinion on whether the deviation/deficiency has an impact on the performance and results of the study or the acceptability of the study.

II. The Fish Short-Term Reproduction Assay

A. **Purpose of the Assay**

The fish-short term reproduction assay (FSTRA) is intended to identify compounds that may have potential to interfere with the normal structure and function of the hypothalamic-pituitary-gonadal (HPG) axis. The reproductive output of a model fish (the fathead minnow, *Pimephales promelas*) exposed to a compound is thought to predict the potential for that compound to interact with HPG components of the endocrine system. The FSTRA is based on the conserved structures and functions of the HPG axis shared by many vertebrate species. It is an important assay because it assesses a variety of endpoints in sexually mature fish that are responsive to interference from estrogenic substances, androgenic substances, anti-androgenic substances and steroidogenesis inhibitors.

The FSTRA is designed to detect changes in spawning, morphology and specific biochemical endpoints that reflect disturbances in the HPG axis. With the exception of VTG induction in males and tubercle formation (a typically male secondary sex characteristic) in females, most of the endpoints in the assay do not unequivocally identify specific cellular mechanisms of action, but do allow inferences to be made with regard to potential agonistic and antagonistic activity and thus provide guidance to determine whether and which additional tests may be warranted. It is intended to be included in a battery of *in vitro* and *in vivo* tests to identify substances with potential to interact with the endocrine system.

B. **Study Design**

The recommended experimental design entails exposing sexually mature male and female fathead minnow to three chemical concentrations, a dilution water control, and a solvent control (if applicable) for 21 days. There are four replicates of each test treatment. Fish density at test initiation is four females and two males per test tank (replicate) for all treatment groups, including controls. The observational endpoints of the test are survival (mortality), fecundity, fertilization success, gonado-somatic index and histology of the gonads, plasma vitellogenin and sex steroid levels, secondary sex characteristics, and other clinical signs. Body weight and length are also recorded at test termination.

III. Evaluation of Study Conduct

This section provides a summary description of the information that would generally be expected to be obtained from a study that had been conducted following the recommendations in the Test Guidelines. As described in this section, the DER reviewer is responsible for summarizing how the study was conducted, the extent to which that is consistent with the Guidelines, and how, if at all, that affected the validity of the study.

This information will factor into the Agency's interpretations of the data contained in the study report. Specific points that are important for the DER to address are highlighted in the individual sections below, as appropriate.

The summary in this section is offered as a general outline to aid in preparation of the DER. The purpose of this section is not to serve as substitute for the Test Guidelines, nor to provide any guidance on how the study should be conducted. Rather, the summary is intended to provide context and examples illustrating to the individual preparing the DER what the DER would be expected to contain.

A. **Test Species**

Fathead minnow (*Pimephales promelas*) is the recommended species. The assay method was validated using reproductively active (4.5-6 months old) groups of this species. If an alternate species is used, the DER should detail the relevant or significant deviations in the method in order to accommodate the alternate species. The DER should include a discussion of whatever evidence was used in the study report for the performance criteria that were used to support the reliability of the test.

B. **Equipment and Supplies**

The list of equipment and supplies from the test guideline is provided only as a non-exhaustive recommendation of what is typically needed to conduct a successful test. If equipment and supplies are used that differ from those identified in the test guideline, it is recommended that the DER identify the differences and state whether and how they may have affected the performance or outcome of the study.

C. **Chemical Testability**

The DER should summarize the results of any tests conducted to evaluate the extent to which the concentrations and stability of the test chemical in the exposure system were verified. The FSTRA is based upon an aqueous exposure protocol whereby test chemical is introduced into the test chambers via a flow-through system. If a successful test is not possible for the chemical using a flow-through test system, a static renewal system may be employed; however, the Corrections and Clarifications document notes that a static-renewal system is not recommended for this assay (EPA 2011). If neither system is capable of accommodating the test chemical even using approved co-solvents, then the default is to not test it using this protocol.

D. **Exposure System**

1. System Description

A flow-through diluter system is preferred (EPA 2011, OECD 2000). The DER should describe the system components, the extent to which they generally comport with

the exposure system description in the guideline, and the extent to which they are capable of maintaining the experimental conditions recommended in the guideline.

2. Water Quality

It is recommended that a description of the source water and chemical analysis results be provided in the DER, in addition to evidence that the water can support normal growth, development, and reproduction of the fathead minnow.

E. Adult Care and Feeding

The test guideline recommends that fish be fed at least twice per day with brine shrimp, or other suitable uncontaminated food, at a rate sufficient to promote active reproduction and maintain body condition. The test guideline further recommends that food be withheld from the fish for 12 hours prior to the day of sampling (at test termination) to simplify histological processing of small fish. More detailed recommendations are found in the test guideline. The DER should summarize conditions during acclimation, pre-exposure, and the definitive test, and state whether these conditions were consistent with the guideline recommendations. In addition, it is helpful to describe in the DER any alternative animal care methods used, along with whether deviations from the recommended methods may have had any significant impact on the study performance or interpretation, using the criteria identified in Table 1 of the test guideline as a guide.

F. Analytical Chemistry

The guideline recommends that test solutions from each replicate tank at each concentration be sampled for analytical chemistry at test initiation (day 0), prior to adding fish. Additionally, the guideline recommends (at a minimum) that weekly analysis be performed on two of the four replicates for every treatment level, including controls. For additional verification of system performance, it is recommended to analyze samples during system preparation and when stock solutions are changed, especially if the volume of the stock solution does not provide adequate amounts of chemical to span the duration of routine sampling periods. The sampling schedule and analyses performed should be provided in the DER.

G. Selection of Test Concentrations

1. Establishing the High Test Concentration

The test guideline recommends that the high test concentration be set based on existing toxicity data for the fathead minnow (*e.g.*, 96-hour LC_{50} value), when available. It is recommended that the high test concentration demonstrate $\geq$ 90% survival in adult fathead minnow, up to the solubility limit of the chemical. If data for the fathead minnow are unavailable, a range-finding test is recommended. The DER should describe the basis

for selecting the highest test concentration included in the study, including a description of any range-finding test(s) conducted and the results.

2. Test Concentration Range

The recommended minimum number of test concentrations is three plus a negative (clean water) control, and a solvent/vehicle control if necessary. The recommended minimum differential between the highest and lowest test concentrations is approximately one order of magnitude [lowest = 0.1X(highest)] (EPA 2011). The minimum recommended separation between individual concentrations is a factor 0.33X and the maximum is a factor of 0.1X (EPA 2011). The DER should identify the range of test concentrations and include any justification provided in the study report, if the test employed a range different from the recommended guidance.

H. Test Procedure

1. Selection and Pre-Exposure

It is recommended that the DER identify the selection and assignment method for fish used in the definitive test, along with the age, weight, and length (if reported) of fish prior to test initiation. The guideline recommends that individual fish selected for the test be within ±20% of the mean weight for male and female fathead minnow, respectively, as determined from a subsample of each sex weighed prior to the pre-exposure phase. Fish are then assigned to replicate tanks (2 males and 4 females each) for the pre-exposure phase, during which time relative reproductive performance is established and any changes in phenotypic (observed) sex ratio are noted. Replicates that meet the recommended minimum criteria for pre-exposure performance (the guideline recommends zero mortality, correct sex ratio, $\geq$15 eggs/female/reproductive day and $\geq$2 spawns in the 7 days prior to the definitive test) are then assigned to randomized blocks for the definitive test. The guideline recommends that replicates be ranked in order of reproductive performance and that then be assigned, in that order, to blocks that will each contain a replicate of each treatment level and controls. If the selection or assignment methods differed from the guideline recommendations, the DER should describe the differences and summarize what impact, if any, the alternative method may have had on the performance or interpretation of the study.

2. Day 0 (Test Initiation)

The definitive test is initiated (day 0) when test item concentrations are established in the test system and spawning groups of fish (replicates), established and assigned during the pre-exposure period, are transferred to the test system. The guideline recommends that survival, reproduction (number of eggs and number of fertilized eggs), secondary sex characteristics, and other clinical signs be recorded daily throughout the test. The DER should identify which observations were made and whether raw (individual) data were provided for the various endpoints.

3. Day 21 (Test Termination)

At test termination, the guideline recommends that blood (plasma) samples be taken from adults and analyzed for vitellogenin (VTG) and for optional sex steroids. Given the limited plasma available from individual fathead minnow, VTG measurements are recommended for all fish, followed by testosterone (T) for male fish and 17β-estradiol (E2) for female fish. If sufficient plasma is available, T measurements are also recommended for female fish, and E2 measurements are recommended for male fish. The guideline recommends that body weight, length, secondary sex characteristics (including nuptial tubercle score) and other clinical signs be recorded for each fish. Gonadal weight, stage, and other histopathological observations indicated in the guideline are recorded. The DER should identify which observations were made and whether raw (individual) data were provided for the various endpoints.

I. Determination of Biological Endpoints

1. Survival, Secondary Sex Characteristics, and Clinical Signs

The DER should indicate whether survival, qualitative observations of secondary sex characteristics, and other clinical signs were recorded on a daily basis. It is recommended that daily survival values be used in calculations of fecundity (below).

2. Body Weight and Length

The guideline recommends that body weight be recorded to the nearest 0.01 g for individual fish at test termination. Individual body weight values are then used to calculate gonado-somatic index (GSI), described in **Section (III)(I)(6)** below; the reviewer should note, however, that body weight is expressed in mg (not grams) for the recommended calculation of GSI. The test guideline also recommends that the length of each fish be recorded to the nearest 0.1 mm at test termination. The DER should state whether body weight and length were recorded at test termination and if the methods (including the level of precision) were consistent with the guideline.

3. Fecundity

The guideline recommends that fecundity be calculated as the number of eggs produced per surviving female per spawn, and that it be measured on the basis of the reproductive day, which is adjusted for survival (mortality). For instance, in a typical 21 day test, if all four females (♀) in a replicate survive, there are 84 reproductive days per replicate (21 days x 4 ♀ = 84 reproductive days). If one female dies 14 days after test initiation, the test consists of 77 reproductive days (3♀ x 21 days + 1♀ x 14 days = 77 reproductive days). The DER should indicate whether the method for calculating fecundity was consistent with the guideline.

4. Fertilization Success

The guideline recommends that fertilization success (fertilization rate) be calculated as the number of embryos divided by the number of eggs, multiplied by 100 (%). Measures are typically made at late cleavage stage (early morning spawn) or tail bud stage embryos (spawned the previous evening). Alternatively, fertilization success can be determined without the aid of a microscope if embryos are counted at the eyed stage (typically 36-48 hours post fertilization). The DER should identify the method used to determine the number of embryos (*e.g.*, which stage was evaluated) and the method used to calculate fertilization success.

5. Nuptial Tubercle Score

The guideline recommends that nuptial tubercles, a secondary sex characteristic typical of male fathead minnow, be mapped and scored for each fish at test termination. The DER should indicate whether tubercle scores were reported for each fish and whether the method for calculating these scores was consistent with the guideline recommendations. Any alternative methods should be briefly described in the DER.

6. Gonado-Somatic Index (GSI)

Gonado-somatic index (GSI) is used in the FSTRA as one metric of reproductive status. The guideline recommends that GSI be calculated as gonad weight (to the nearest 0.1 mg) divided by body weight (mg), multiplied by 100 (%). The DER should state whether GSI was reported and whether the calculation method was consistent with the guideline recommendations. Any alternative methods should be described briefly in the DER.

7. Gonadal Staging and Histopathology

Appendix E of the test guideline provides recommendations for determining the developmental stage of fathead minnow gonads and for evaluating diagnostic histopathology criteria for male and female fish, respectively. The gonadal stage values recommended in the guideline are J=juvenile, 0=undeveloped, 1=early spermatogenic, 2=mid-spermatogenic, 3=late spermatogenic, 4=spent, UTS=unable to stage for male fathead minnow and J=juvenile, 0=undeveloped, 1=early development, 2=mid-development, 3=late development, 4=late development/hydrated, 5=post-ovulatory, and UTS=unable to stage for female fathead minnow. In addition, the guideline identifies primary and secondary diagnostic criteria for histopathology. For males, the recommended primary diagnostic criteria include increased proportion of spermatogonia, presence of testis-ova, increased testicular degeneration, and interstitial (Leydig) cell hypertrophy/hyperplasia; and the recommended secondary diagnostic criteria include decreased proportion of spermatogonia, increased vascular or interstitial proteinaceous fluid, asynchronous gonad development, altered proportions of spermatocytes or spermatids, and granulomatous inflammation. For females, the recommended primary

diagnostic criteria include increased oocyte atresia, perifollicular cell hyperplasia/hypertrophy, and decreased yolk formation; and the recommended secondary diagnostic criteria include interstitial fibrosis, egg debris in the oviduct, granulomatous inflammation, and decreased post-ovulatory follicles. It is recommended in the test guideline that each of the primary and secondary observations be scored using the following severity grades: 0=not remarkable, 1=minimal, 2=mild, 3=moderate, and 4=severe. Further, the guideline states that severity grades may be assigned either based on comparison to the negative control or based on the pathologist's determination of normal histology. The DER should summarize the pathologist's method for determining gonadal stage; the DER should also identify which histopathology criteria and additional characteristics, if applicable, were evaluated by the pathologist, along with the associated severity grading scale(s) and basis for comparison (to negative controls or to normal histology).

8. Vitellogenin (VTG)

The guideline recommends that plasma vitellogenin (VTG) levels be quantified for samples from individual fish at test termination. One method for quantifying VTG is to perform an enzyme-linked immunosorbent assay (ELISA), using monoclonal or polyclonal antibodies and purified vitellogenin protein isolated from the fathead minnow. The DER should state whether VTG was analyzed for male and female fish and should summarize the methods used (*e.g.*, ELISA or alternative method). It is recommended that the DER include a brief description of any proficiency and quality control measures reported by the study author(s). If an alternative method was used, the DER should indicate whether and how it may have impacted the assay performance or interpretation of results.

9. Sex Steroids (Testosterone and 17β-estradiol)

If enough tissue (plasma) is available after samples are taken for VTG, the test guideline recommends determining plasma sex steroid concentrations from the blood samples taken at test termination. The guideline recommends that testosterone (T) measurements take priority for samples from male fish and that 17β-estradiol (E2) measurements take priority for samples from female fish. Thereafter, if enough tissue remains, T measurements for female fish and E2 measurements for male fish can also provide useful information. The DER should identify whether and which sex steroid concentrations were determined for male and female fish, respectively. The DER should also summarize the methods used (*e.g.*, ELISA, radio-immunoassay, or an alternative method) and describe any proficiency and quality control measures used to support the validity of the method.

10. Specimens Archival

It is recommended that any archival practices, *e.g.*, for histopathology slides, digital images, and tissue samples, be documented by the reviewer in the DER.

11. Data Reporting and Completeness

The DER should document whether data were collected using electronic or manual systems which conform to good laboratory practices (GLP). The reviewer should confirm whether the study report contains the information identified for reporting in the test guideline, including but not limited to the endpoints described in the foregoing sections of this SEP.

IV. Study Interpretation

This section of the DER is intended to address the interpretation of the study results and any conclusions regarding the acceptability of the study. As part of this evaluation, the DER should include a discussion of how well the study conforms to specific validity and performance criteria identified in the test guideline; the analysis of validity and performance criteria would draw, in part, upon the evaluation of study conduct described in the previous section. This is intended to clarify the reviewer's conclusions regarding whether, to what extent, and how any deviation(s) affect the interpretation or acceptability of the study. **Section (IV)(D)** provides specific recommendations as to how the biological data obtained from the FSTRA may be interpreted by the reviewer and reported in the DER.

The following sections are based on a summary of the information generally expected to be obtained by a study that was conducted following the Test Guidelines, and that would generally be relevant to interpreting the results of the FSTRA. This summary is provided as a general outline to aid the reviewer in preparing a DER and not as a substitute for the Test Guidelines, nor as guidance on how to conduct the assay.

A. Test Validity Criteria

The DER should note whether the assay met the validity criteria identified in the OCSPP 890.1350 test guideline, based on the data and other information provided in the study report. The validity criteria are summarized below in **Table 1**. The validity criteria are based upon the guideline-recommended experimental design and the associated statistical power of the FSTRA.

When met, the validity criteria are used to support the scientific soundness of conclusions drawn from the test results. These validity criteria provide an indication of the general performance of the FSTRA under test conditions; *e.g.*, control survival that does not meet the reference value ($\geq$90%) may be an indication of underlying problems associated with husbandry, environmental conditions, or the particular strain of test organisms used in the study. Therefore, the extent to which the validity criteria are met would be relevant to any conclusions drawn from the study.

The DER should include a discussion of any rationale provided in the study report for any validity criteria that were not met. This includes identification of whether, to what extent, and in what way failure to meet the criteria has had an impact on the quality or acceptability of the study.

Table 1: Validity Criteria for the Fish Short-Term Reproduction Assay.

Criteria	Reference Value(s)
Survival (in controls)	≥ 90%
Reproduction (in controls, each replicate) Spawning Fecundity Fertilization success	 at least every four days, OR ≥ approximately 15 eggs/female/reproductive day/replicate; AND ≥ 95%
Temperature (mean)	25±1°C
Dissolved oxygen	≥ 4.9 mg/L (≥ 60% saturation)
Measured test concentration (each test item treatment, all replicates)	CV < 20% over 21 days

Abbreviations: CV Coefficient of variation.

B. **Performance Criteria**

Performance criteria applicable to the guideline are summarized in **Table 2** below. (Criteria that are identified in the guideline as both validity criteria and as performance criteria are not repeated in this section.) The DER should include enough information to determine whether the assay met the relevant criteria.

Where deviations from the guideline are reported the performance criteria may be used to demonstrate that such deviations had a minor impact on study outcome. A study that fails to meet one or more performance criteria may still provide useful information and is not necessarily rejected. The DER includes a discussion of any rationale provided in the study report for any performance criteria that were not met. That discussion should also identify whether, to what extent, and in what way failure to meet the specified performance criteria had an impact on the quality or interpretation of the study. For additional guidance on the interpretation of performance criteria related to analytical chemistry and general water quality parameters, the reviewer may consult OCSPP Guideline 850.1000, Special Considerations for Conducting Aquatic Laboratory Studies.

Table 2: Performance Criteria for the Fish Short-Term Reproduction Assay.

Criteria	Reference Value(s)
Dilution water quality	
Total alkalinity	> 20 mg/L as $CaCO_3$
Total organic carbon	$\leq$ 2 mg/L
Unionized ammonia	$\leq$ 1 µg/L
Residual chlorine	< 10 µg/L
pH	6.5-9.0
Water temperature	25±1°C ($\leq$1°C difference between replicates at any time during the test)

C. General Analysis

1. Statistical Analyses

Where appropriate, formal statistical analysis is used to identify significant differences. Recommended statistical procedures for analyzing survival, fecundity and fertility, gonado-somatic index (GSI), vitellogenin (VTG), sex steroids (if measured), nuptial tubercle score, weight, and length are summarized in the **Appendix**. As recommended by the guideline, all statistically significant differences would be reported by endpoint and by concentration. The direction (*e.g.*, increase or decrease) and magnitude (*e.g.*, percent change as compared to control) are stated, where relevant. The level of significance (*e.g.*, $\alpha = 0.05$) would also typically be reported for each test when there is a statistical significance; to aid in interpretation, it is helpful to report actual p-values for all endpoints evaluated. Specific guidance for the statistical analysis of histopathology findings is not provided, but summary statistics for severity and incidence of observations should be reported in the DER.

a. Statistical Tests Employed

The DER should identify the extent to which the statistical methods used were consistent with those recommended in the test guideline. The DER should include the statistical tests (and software, as appropriate) used, along with any rationale provided for the use of either parametric or nonparametric tests. Any data transformations would also be reported.

The extent and specific nature of statistical verification by the reviewer may vary. Typically, the reviewer confirms the accuracy of the statistical analyses by recalculating summary statistics and pertinent statistical tests for endpoints. This verification is facilitated if individual data for all results, including negative and solvent controls (if applicable), are submitted with the study report in electronic format (*e.g.*, spreadsheet files).

b. Test Using Solvents

If a solvent was used, confirmation of comparisons between solvent controls and clean water controls is typically performed to determine potential solvent effects on the organisms. All control results that have been provided are included in the DER. Evidence of a solvent effect will be considered in determining the utility of a study. The reviewer should consult existing guidance on this topic (*e.g.*, OCSPP 850.1000, EPA 2008). In addition, one or both of the following analyses is/are typically performed:

- Clean water control versus treatment concentrations: Statistical comparison of responses in the test concentrations in relation to the negative (dilution water) control is recommended for all tests, and all control results are included in the DER. Unless stated in other policy or guidance documents, this method is the preferred method for identifying treatment-related effects in the FSTRA and also for the reviewer to generally employ in generating the DER.

- Solvent control versus treatment concentrations: Statistical comparison of responses in the test concentrations in relation to the solvent control is another method of evaluation. All control results should be included in the data report.

c. Outliers

The DER should also identify the method used to identify any outliers, if applicable. If outliers were excluded from the statistical analysis in the DER, the biological and/or statistical justification should be described.

2. Trends

The DER should report all trends, whether positive or negative. A discussion of the significance of the trends and implications for the interpretation of the test is also suggested.

3. Histological Findings

The DER should report gonadal stage and other histological findings by diagnostic criterion, severity grades, and opinion of the pathologist and reviewer. As indicated above, summary statistics are to be presented and verified by the reviewer.

D. Endpoint Interpretation

The FSTRA as presented is intended to serve in a screening capacity to provide an indication of potential endocrine activity, but not to confirm a specific mechanism, mode of action, or adverse effect. Therefore, a significant (especially statistically significant, when endpoints are amenable to statistical analysis) effect in one or more of the core endpoints of this assay (fecundity, fertilization success, nuptial tubercle score, GSI,

gonadal histopathology, or VTG) may be indicative of a potential of the test material to disturb the HPG axis of fishes. The full suite of endpoints is helpful to provide a comprehensive assessment of the potential of the chemical to interact with the HPG axis in a representative vertebrate.

1. Survival, Secondary Sex Characteristics, and Clinical Signs

Survival (mortality) and clinical signs of toxicity, including behavioral effects, may be used to determine whether a treatment demonstrates overt toxicity. Changes in behavior may include but are not limited to hyperventilation, uncoordinated swimming, unusual quiescence, loss of equilibrium or altered feeding behavior. Certain behavioral changes, such as a decline in territorial behavior, can also be indicative of exposure to endocrine active compounds when observed in treated fish but not (or to a lesser extent) in controls.

Observations of secondary sex characteristics may include (but are not limited to) body color (light or dark), coloration patterns (such as the presence or absence of vertical bands), body shape (*i.e.* shape of the head and pectoral region, abdomen distension), and specialized secondary characteristics (*e.g.*, changes in the size of the dorsal nape pad). It is recommended that nuptial tubercle score, another specialized secondary sex characteristic, be quantitatively evaluated for specimens at test termination; this endpoint is described separately in **Section (IV)(D)(4)** below.

The DER should summarize any effects on survival and report the incidence of secondary sex observations and clinical signs, by type, for each sex. If appropriate, effects on survival should be analyzed statistically, and methods and results should be reported in the DER.

2. Body Weight and Length

Body weight and length are included as supporting measurements that are not specific to (diagnostic of) reproductive or HPG effects. Effects on body weight and length should be interpreted with consideration to other endpoints in the FSTRA, as growth effects may result either from generalized toxicity or from endocrine-mediated effects. The DER should report summary data (*e.g.*, replicate means and standard deviations) for body weight and length in male and female fish, respectively, at test termination. Results of any statistical analysis by the study author(s) and statistical verification by the reviewer should also be described in the DER.

3. Fecundity and Fertilization Success

The DER should report summary data (*e.g.*, replicate means and standard deviations) for fecundity and fertilization success; the recommended calculations for these endpoints are described in Section III of this SEP. Results of any statistical analysis

by the study author(s) and statistical verification by the reviewer should also be described in the DER.

4. Nuptial Tubercle Score

The DER should report replicate median nuptial tubercle scores for male and female fish in each treatment and controls, including the solvent control, if applicable. Results of any statistical analysis by the study author(s) and statistical verification by the reviewer should also be described in the DER.

5. Gonado-Somatic Index (GSI)

Control GSI values in fathead minnow are typically 1-2% for males and 8-13% for females, with high degrees of individual variation (*e.g.*, Ankley *et al*. 2001). The DER should report summary data (*e.g.*, replicate means and standard deviations) for GSI in male and female fish, respectively. Results of any statistical analysis by the study author(s) and statistical verification by the reviewer should also be described in the DER.

6. Gonadal Staging and Histopathology

Specific alterations of gonad histology may be indicative of particular modes of action of endocrine active compounds. Gonads can be assessed using routine histological protocols; specific guidelines for post-mortem and histological examination are provided in the test guideline and summarized in **Section III** of this SEP. The DER should include summary statistics for gonadal stage (replicate median developmental stage values) and incidence values for each diagnostic criterion, by severity grade, for male and female fish, respectively. The determination of an effect associated with exposure to a chemical may be heavily weighted by the expert opinion of a qualified pathologist, as presented in the histopathology description of the study report and verified by the reviewer.

7. Vitellogenin (VTG) and Sex Steroids

The DER should report summary data (*e.g.*, replicate means and standard deviations) for vitellogenin and sex steroids (if measured) in male and female fish, respectively, preferably expressed as ng/mL plasma. Results of any statistical analysis by the study author(s) and statistical verification by the reviewer should also be described in the DER.

E. Special Data Analysis Considerations

1. Use of Compromised Treatment Levels

Several factors are typically considered when determining whether a replicate or entire treatment demonstrates overt toxicity and should be removed from the analysis. Overt toxicity is generally defined in the FSTRA test guideline as greater than two

mortalities in any replicate, which can only be explained by toxicity and not technical error. Other signs of overt toxicity include hemorrhage, abnormal behaviors, abnormal swimming patterns, anorexia, and any other clinical signs of disease. For sublethal signs of toxicity, qualitative evaluations may be necessary and would be made in reference to the negative (clean water) control group.

2. Solvent Controls

The use of a solvent is typically only considered as a last resort, after all other chemical delivery options have been determined to be inappropriate, because it has the potential to interfere with the test results. If a solvent is used, then a clean water control is typically run in concert. At the termination of the test, an evaluation of the potential effects of the solvent is performed. This is typically done through a statistical comparison of the solvent control group and the clean water control group. If significant differences are detected between the clean water control and solvent control groups, the guideline recommends that best professional judgment be used to determine if the validity of the test has been compromised. If there are no significant differences between the clean water control and solvent control for any of the measured response variables, it is recommended that the study endpoints be determined based on comparison to the clean water control unless other guidance or policy recommends otherwise.

V. Data Evaluation Record

It is recommended that the submitted study be reviewed according to the principles in the previous sections of this SEP. The review is then documented in the Data Evaluation Record (DER). A template that provides additional guidance to the reviewer for preparation of the DER is available. Generally, the DER will include a cover sheet, executive summary and results synopsis, quality assurance elements, description of the material and methods used, summary of the study author's reported results and analysis, and the reviewer's interpretation of results and discussion. The DER should typically identify deviations from the guideline-recommended methods and validity and performance criteria, along with a discussion of their significance. The DER should generally identify the effective concentrations associated with each endpoint in the FSTRA. Finally, the DER will include a conclusion as to whether the test item was potentially active in the FSTRA HPG axis in the submitted study [see **Section (IV)(D) above**], and whether the study satisfied the Test Order for an Fish Short-Term Reproduction Assay using OCSPP Guideline 850.1350.

VI. References

Ankley, G.T., K.J. Jensen, M.D. Kahl, J.J. Korte, and E.A. Makynen. (2001). Description and evaluation of a short-term reproduction test with the fathead minnow (*Pimephales promelas*). *Environmental Toxicology and Chemistry* 20: 1276 – 1290.

Organization for Economic Cooperation and Development (OECD). (2006). Current approaches in the statistical analysis of ecotoxicity data: A guidance to application. OECD Environmental Health and Safety Publications, Series on Testing and Assessment No. 54. Paris, France.

Organization for Economic Cooperation and Development (OECD). (2000). Guidance document on aquatic toxicity testing of difficult substances and mixtures. OECD Environmental Health and Safety Publications, Series on Testing and Assessment No. 23.

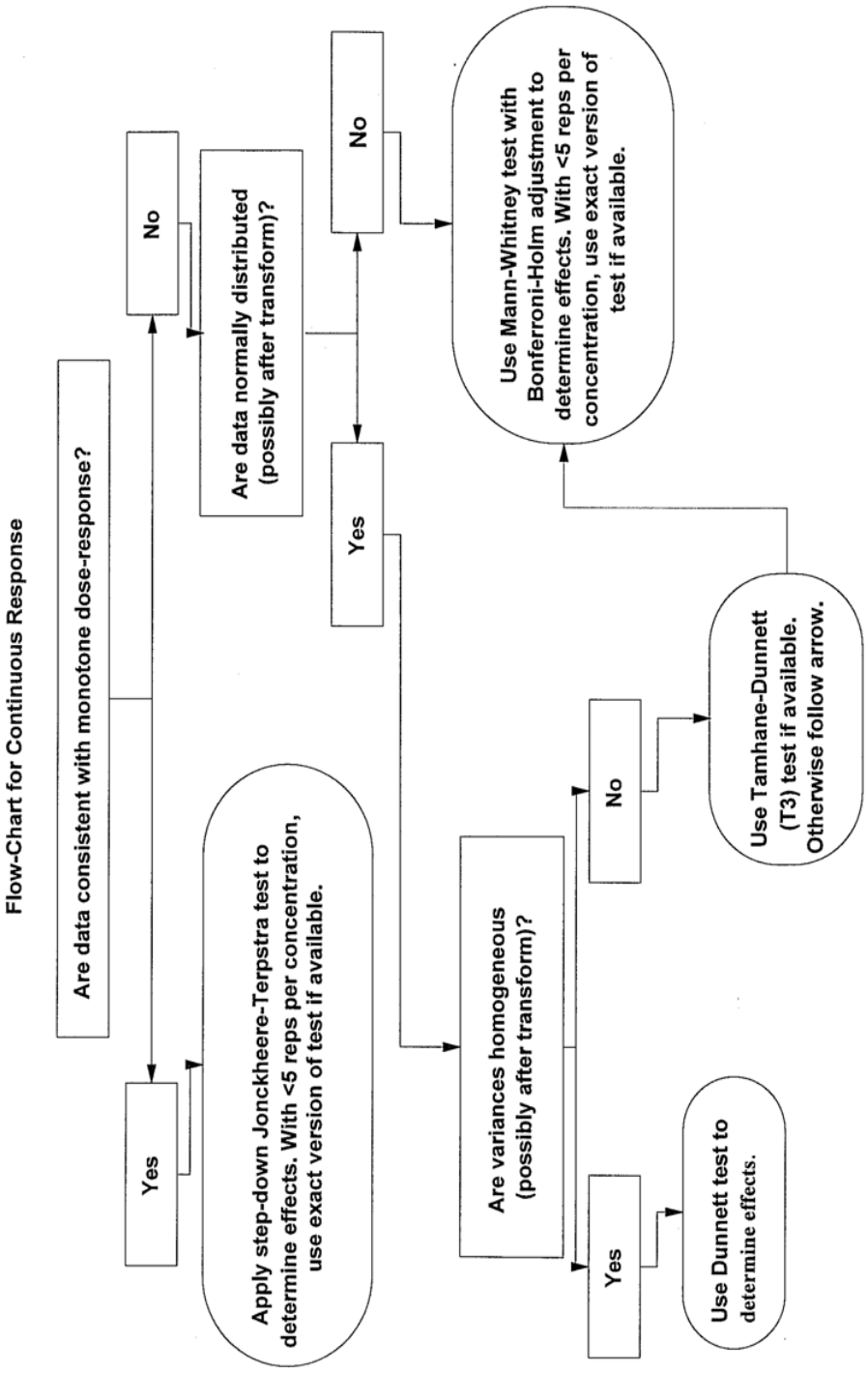
U.S. Environmental Protection Agency (EPA). (2011). Corrections and Clarifications on Technical Aspects of the Test Guidelines for the Endocrine Disruptor Screening Program Tier 1 Assays (OCSPP Test Guideline Series 890). March 3, 2011. Office of Chemical Safety and Pollution Prevention (OCSPP), Washington, D.C. (http://www.epa.gov/endo/pubs/assayvalidation/clarificationdoc.pdf).

U.S. Environmental Protection Agency (EPA). (2008). Guidance for the use of dilution-water (negative) and solvent controls in statistical data analysis for guideline aquatic toxicology studies. September 25, 2008. Memo from Statistics Workgroup and Aquatic Biology Technical Team to Donald Brady, Director, Environmental Fate and Effects Division. Office of Pesticide Programs, Office of Chemical Safety and Pollution Prevention (OCSPP), Washington, D.C.

Appendix: Recommendations for Statistical Analyses Based on OECD 2006

The statistical analyses recommended in the FSTRA test guideline (OCSPP 890.1350) are described in more detail in the document, Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application (OECD 2006). For all continuous quantitative endpoints (fecundity, fertility, GSI, VTG, sex steroids, weight, and length) that are consistent with a monotonic dose-response, it is recommended that the Jonckheere-Terpstra test be applied in step-down manner to establish a significant treatment effect. For continuous endpoints that are not consistent with a monotone dose-response, it is recommended that the data be assessed for normality (preferably using the Shapiro-Wilk or Anderson-Darling test) and variance homogeneity (preferably using the Levene test). Both tests are performed on the residuals from an ANOVA. Expert judgment can be used in lieu of these formal tests for normality and variance homogeneity, though formal tests are preferred. Where non-normality or variance heterogeneity is found, a normalizing, variance stabilizing transformation is recommended. If the data (perhaps after a transformation) are normally distributed with homogeneous variance, it is possible to determine a significant treatment effect using Dunnett's test. If the data (perhaps after a transformation) are normally distributed with heterogeneous variance, a significant treatment effect may be determined from the Tamhane-Dunnett or T3 test or from the Mann-Whitney-Wilcoxon U test. Where no normalizing transformation can be found, it is possible to determine a significant treatment effect from the Mann-Whitney-Wilcoxon U test using a Bonferroni-Holm adjustment to the p-values. The Dunnett's test is applied independently of any ANOVA F-test and the Mann-Whitney test is applied independently of any overall Kruskal-Wallis test.

Significant mortality is not expected but, if it occurs, can be assessed from the step-down Cochran-Armitage test, where the data are consistent with dose-response monotonicity, and otherwise from Fisher's Exact test with a Bonferroni-Holm adjustment. A significant treatment effect for nuptial tubercle score may be determined from the step-down application of the Jonckheere-Terpstra test applied to the replicate medians. Alternatively, and preferably, it is possible to use the multi-quantal Jonckheere test for effect determination, as it takes into account changes to the distribution profile. The appropriate unit of analysis is the replicate so the data consist of replicate medians if the Jonckheere-Terpstra or Mann-Whitney U test is used, or the replicate means if Dunnett's test is used. Dose-response monotonicity can be assessed visually from the replicate and treatment means or medians or from formal tests. With fewer than five replicates per treatment or control (which would be expected if the FSTRA were performed exactly according to the guideline), the exact permutation versions of the Jonckheere-Terpstra and Mann-Whitney tests is recommended if available. It is typically recommended that the statistical significance of effects be judged at the 0.05 significance level, unless otherwise specified in the DER.


Flow-Chart for Continuous Response
Are data consistent with monotone dose-response?
Yes
No
Apply step-down Jonckheere-Terpstra test to determine effects. With <5 reps per concentration, use exact version of test if available.
Are data normally distributed (possibly after transform)?
Yes
No
Use Mann-Whitney test with Bonferroni-Holm adjustment to determine effects. With <5 reps per concentration, use exact version of test if available.
Are variances homogeneous (possibly after transform)?
Yes
No
Use Dunnett test to determine effects.
Use Tamhane-Dunnett (T3) test if available. Otherwise follow arrow.