Pubertal Development and Thyroid Function in Intact Juvenile/Peripubertal Female Rats Assay

OCSPP Guideline 890.1450

Standard Evaluation Procedure (SEP)

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

August 2011

Table of Contents

I. INTRODUCTION

A. Use of the Standard Evaluation Procedure

This Standard Evaluation Procedure (SEP) was developed by the U.S. Environmental Protection Agency (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). The SEPs provide general guidance and are not binding on either EPA or any outside parties. The use of language such as "will," "is," "may," "can" or "should" in these documents 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 SEP provides guidance on how EPA generally intends to review studies conducted using the OCSPP Guideline 890.1450 for the Female Pubertal Assay that are submitted to support requirements imposed under the EPA's EDSP. The product of the review will be a Data Evaluation Record (DER) that reflects how well the study was performed and conforms to the Guideline and provides the appropriate conclusions supported by the data. The DER is intended to include, for example, a list of any significant deviations from the protocol as well as their potential impacts, a list of significant information missing from the study report, and any other information about the performance of the study that affects interpretation of the data within the context of the EDSP.

The DER is intended to contain adequate information to provide the EPA with the ability to determine whether the study was performed according to the Guideline. The objective of EDSP Tier 1 assays is to characterize the potential of a chemical to interact with the endocrine system.

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 named 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 — or that 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 FEMALE PUBERTAL ASSAY

A. Purpose of the Assay

The purpose of the female pubertal assay is to help identify chemicals that have the potential to interact with the endocrine system, by identifying effects on pubertal development and thyroid function in the intact juvenile/peripubertal female rat. This assay is capable of detecting not only anti-thyroid, estrogenic or anti-estrogenic chemicals, but also agents that alter pubertal development through mechanisms that induce changes in luteinizing hormone, follicle

stimulating hormone, prolactin or growth hormone levels or via alterations in hypothalamic function.

B. Background

Endocrine disrupting xenobiotics have the potential to interact with hormone systems. The EDSP reflects a two-tiered approach to implement the statutory testing requirements of FFDCA section 408(p) (21 U.S.C. 346a) and SDWA (42 U.S.C. 300j-17). In general, EPA intends to use the data collected under the EDSP, along with other scientifically relevant information (OSRI), to determine if a pesticide, chemical, or other substance, may pose a potential hazard to human health or the environment due to disruption of the endocrine system. EPA has developed the OCSPP 890 Series Guidelines as a screening battery (Tier 1 screening) to identify these substances.

C. Study Design

In the general design of a Pubertal Female Rat Assay, groups of juvenile female rats are exposed to the test substance daily by oral gavage from post-natal day (PND) 22 through PND 42. This duration of treatment is necessary for the detection of pubertal delay and anti-thyroid effects. The rats are either treated with vehicle only, or the test substance in the vehicle at two dose levels. The evaluation includes observations to detect anti-thyroid, estrogenic or anti-estrogenic chemicals or agents which alter pubertal development.

D. Endpoints Evaluated

The following endpoints are evaluated in the Female Pubertal Rat Assay:

Growth	Daily body weight
Vaginal opening (VO)	Age and weight at VO
Organ weights	Thyroid (after fixation) Liver Kidneys (paired) Pituitary Adrenals (paired)
Histology	Uterus Ovary Thyroid (colloid area and follicular cell height) Kidney
Hormones	Serum thyroxine (T_4), total Serum thyroid stimulating hormone (TSH)
Estrous cyclicity	Age at first vaginal estrus after vaginal opening Length of cycle Percent of animals cycling Percent of animals cycling regularly
Clinical (serum) chemistry	Standard blood panel, including creatinine and blood urea nitrogen (BUN)

III. EVALUATION OF STUDY CONDUCT

A. Test Compound

The purity of the test compound used in the study, its source, and Lot No. and/or Batch No. are reported in the DER, along with a statement that the certificate of analysis was provided. The stability of the test chemical in the vehicle should be reported in the DER for concentrations used in the study, and the storage temperature used for the stability analyses should be specified. For test substances dosed as suspensions, the homogeneity of the test chemical suspensions should also be reported in the DER. The DER should provide the range of values expressed as percent of nominal for the analysis of test chemical solutions/suspensions.

EPA recommends that the dosing vehicle used for the study not have any potential intrinsic toxicity (e.g., acetone, DMSO) in order to avoid confounding the study results. Other solvents, such as water or carboxymethylcellulose, may be used where appropriate. Use of an intermediate solvent would not be expected to affect the validity of the data significantly, provided the concentration is kept at low concentrations (e.g., 1% or below) and is used across all test groups including control. EPA recommends that the study design include the use of a solvent or vehicle control group.

Gentle warming may be effectively used to assist solubilization, but it is recommended that the solution not be administered warm and be checked to make sure that precipitation did not occur upon cooling. If the test substance is not soluble in any of the conventional solvents, EPA recommends that it be administered as a suspension. It is important that the dosing solution or suspension be well-mixed to keep the chemical well-distributed prior to and throughout dosing, and care must be taken to ensure that the particle size of insoluble substances does not interfere with delivery of the full dose through the gavage tube or needle tip.

B. Test Animals

Sprague-Dawley rats are the preferred strain for this assay until a more-appropriate strain (or set of strains) is identified and associated performance criteria are developed. Results similar to those from Sprague-Dawley rats have been produced using Wistar and Long-Evans rats in this assay or relevant modifications of this assay, suggesting that strain is not the major determinant of sensitivity in this assay. EPA recommends that the DER include: the species and strain used and the rationale for this choice; the source and supplier of the animals; and the number and age of the animals at receipt, dosing initiation, and necropsy.

It is recommended that juvenile female rats be derived from individually housed pregnant females that were either bred in-house or purchased from a supplier as “timed pregnant” dams. EPA recommends that dams obtained and transported from an external supplier not be used in the same study as dams bred in-house. There are several reasons that the Agency strongly recommends that pregnant dams rather than pups of a specified age be used as the starting point for the pubertal assays. First, it is important to minimize genetic effects (that is, litter effects) on the endpoints of interest where possible. It is important to know which pups are from the same dam in order to be able to randomize distribution of siblings across treatment groups. Second, it is important to cull to 8 to 10 pups per litter within 3 or 4 days after birth. This helps to minimize variability in body weights across a litter, and thus minimize variability in day or

vaginal opening and in other endpoints such as organ weights that may be related to body weight. Third, it is important to know the day of birth of each pup accurately. Knowing the day of birth accurately is critical for accurate determination of age at vaginal opening and keeping the coefficients of variation (CVs) low. Keeping the CVs as low as possible is important for maximizing the sensitivity of the pubertal endpoints.

Thus, if pups of a specified age are ordered rather than pregnant dams, the Agency recommends that submitters document that all of the necessary steps have still been taken to allow randomization of litters across treatment groups, and that standardization of litters (including exclusion of litters with fewer than 8 total pups per litter as well as litters not delivered by GD 23, reduction of litter size to 8-10 pups per litter between post-natal day 3 and 5 and not allowing cross-fostering) has been done.

It is also recommended that all dams be pregnant for the first time and timed to deliver on the same day. If purchased from a supplier, EPA recommends that all dams be on the same gestation day (GD), but does not believe that whether the day is GD 7, 8, 9, or 10 at the time of arrival at the performing laboratory (where GD 0 is defined as the day the dams are found to be sperm positive) is likely to affect study results. EPA recommends that dams be allowed to deliver their pups naturally.

The Guideline recommends pup body weights be monitored weekly and any unthrifty litters or runted pups not be included in the study. The Guideline also recommends littermates not be used in the same experimental group; and at least 15 female pups be used per treatment group. Typically, the pups are weaned on PND 21, and the female pups assigned to treatment groups such that the mean body weights and variances for all groups are similar. It is imperative that treatments be initiated no later than PND 22. Initiating treatment later than PND 22 can result in failure of the study, as the onset of puberty (i.e., vaginal opening) in the control female rats will begin within a few days.

C. Animal Husbandry

The Guideline recommends that the study be conducted in facilities accredited by the Association for Assessment and Accreditation of Laboratory Animal Care International (AAALAC) if in the U.S., or the applicable national or international accreditation authority if outside the U.S. Rats are housed in cages with heat-treated (to eliminate resins that induce liver enzymes) laboratory-grade wood shavings (other than cedar) as bedding. The Guideline recommends corn cob bedding not be used due to its potential to disrupt endocrine activity (Markaverich *et al*., 2002); and wire-mesh-bottomed caging not be used due to the potential for pup loss.

The Guideline recommends animals be maintained on a balanced laboratory diet that has a genistein-equivalent content of genistein plus daidzein (aglycone forms) less than or equal to 300 µg/g, and the same batch of feed be used for treated and control groups at all times.

Deionized water is recommended in the Test Guideline. Other acceptable sources of water include double-distilled water and charcoal-filtered water. Other sources may be acceptable; however, the presence of soluble organic chemical contaminants such as natural or artificial hormones have the potential to introduce variability into (and potentially compromise) the

results. If an alternative source of water has been used, the Agency recommends that the laboratory document that such contaminants have been removed from the drinking water. The Test Guideline recommends that water not be supplied using polycarbonate supply equipment.

D. Experimental Design

The Guideline specifies use of a randomized complete block design (time-separated necropsy is the blocking factor) with at least 15 female rats in each treatment group. The treatment groups include: (1) the vehicle-treated and (2) xenobiotic-treated with at least two dose levels. The Guideline recommends that the highest dose level be at or just below the Maximum Tolerated Dose (MTD) but need not exceed the Limit Dose (1000 mg/kg/day); however, typically, the Agency also considers the toxicity profile of the chemical (i.e., cholinesterase inhibition, target organ toxicity, etc.) in dose selection. The second dose level is spaced to produce a lesser degree of toxicity relative to the high dose unless justification is provided for testing at a different level. The DER should contain the rationale for the selection of doses. If necessary, it is possible to conduct the study in time-separated blocks rather than at one time without affecting the validity of the results. In this case, each block is recommended to contain all treatment groups (i.e., control, vehicle control, etc.) and be balanced with respect to numbers of animals and body weight at weaning.

E. Dosing

The Guideline recommends that animals be weighed daily, prior to treatment, and the body weight recorded. Consistent with good scientific practice, clinical observations are recorded daily. Endpoint measurements (organ weights, hormone levels, histology, etc.) from animals which were found dead or euthanized *in extremis* during the study are not included in the summary statistics.

The Guideline specifies treatments be administered daily by oral gavage from PND 22 through PND 42. Although the duration of treatment is longer than necessary to detect estrogenic chemicals, this duration is needed to detect delays in puberty and anti-thyroid effects. The Agency recommends test chemicals be administered daily by gavage at a dose volume of 2.5 to 5.0 mL vehicle/kg body weight at 0700-0900 daily. The Guideline recommends dosing equipment not utilize any plastic or rubber to avoid the potential for absorption by or leaching of substances. The treatments are typically administered on a mg/kg body weight basis, using the current day's weight, and volume of the dose administered recorded each day.

In the absence of other clinical signs that would normally lead to removal of an animal from the study, failure to gain weight at the same rate as controls is generally considered not to be a reason to remove a treated animal during the course of the study. However, it is recognized that severe failure to grow may be a reason to disqualify an animal even in the absence of other signs of toxicity. These types of clinical sign are to be reported in the DER.

F. Vaginal Opening

Beginning on PND 22, females are examined daily for vaginal opening (VO). The appearance of a small "pinhole," a vaginal thread, and complete VO are recorded on the days they are observed. The day of complete VO is the endpoint used in the analysis for the age at vaginal opening; a "pin hole" or thread does *not* represent complete vaginal opening, even

though the Guideline recommends it be recorded when observed. However, if any animal within any treatment group shows incomplete opening (such as persistent threads or a "pin hole") for greater than three days, it is recommended a separate analysis is typically conducted using the ages at which incomplete opening was first observed. Documentation of a vaginal thread, even if vaginal opening otherwise appears complete, is important. It is also critical that "initiation" of vaginal opening be recorded in order to detect accurate timing of initiation of VO. It is preferred but not critical that vaginal opening observations be taken after the daily dosing. Whether collected before or after dosing, it is critical that the vaginal opening observations be collected at approximately the same time each day in order to detect accurate timing of initiation of VO. The age and body weight at vaginal opening is generally reported in the DER for each treatment group.

It is critical that, for each animal, the VO observation be recorded for the day immediately prior to the day on which VO begins in order to detect accurate timing of initiation of VO. Laboratories may choose not to begin monitoring on PND 22, but missing the day on which VO begins for each animal will be considered a serious deficiency in the study because the sensitivity of this endpoint is dependent on the accurate determination of the day of VO (or in certain cases, day of initiation of VO if the process is not complete in one day). If the day of VO is expected to be different for control animals from what is noted in the Guideline, EPA recommends that appropriate documentation be provided to the Agency to support a modification in study design to begin observations later than recommended. Note that consideration should be given to the possibility that the test chemical may accelerate VO; consequently, observations for these endpoints should begin substantially before the age at which control animals are expected to reach these endpoints. The objective of the VO endpoint is to determine quantitatively the difference, if any, in age at VO between treated groups and controls, not merely to determine that an acceleration or delay has occurred.

G. Estrous Cyclicity

Beginning on the day of vaginal opening, through and including the day of necropsy, daily vaginal smears are obtained and evaluated for the presence of leukocytes, nucleated epithelial cells, or cornified epithelial cells (Cooper *et al*., 1993). The vaginal smears are classified as diestrus (predominance of leukocytes mixed with some cornified epithelial cells), proestrus (predominance of clumps of round, nucleated epithelial cells), or estrus (predominance of cornified epithelial cells). Metestrus is classified as an early part of diestrus rather than a late part of estrus. The age at first vaginal estrus is reported (Goldman *et al*., 2007) in the DER. A typical cycle consists of two or three days of diestrus, one day of proestrus, and one or two days of estrus. It is preferred but not critical that estrous cycle observations be taken after the daily dosing. Whether collected before or after dosing, the Guideline recommends estrous cycle observations be collected at approximately the same time each day.

At the end of the study, the overall pattern of each female is characterized as regularly cycling (having recurring 4- to 5-day cycles), irregularly cycling (having cycles with a period of diestrus longer than 3 days or a period of cornification longer than 2 days), or not cycling (having prolonged periods of either vaginal cornification or leukocytic smears). In cases where there are too few days between vaginal opening and the end of the study to observe more than one cycle, classification will be based on the available data with the default assumption that

animals are cycling regularly if the partial data fit the definition, and are irregularly cycling if the study ends without being able to distinguish between irregular cycling and not cycling.

H. Terminal Procedures

1. Necropsy

The Guideline recommends females be sacrificed on PND 42 beginning 2 hours after the final dose. If necessary, one half of the females can be sacrificed on PND 42 and the remaining females on PND 43, without affecting the validity of the study results, as long as the animals in each treatment group (including controls) are equally dispersed between the two necropsy days. The Test Guideline recommends that animals sacrificed on PND 43 be dosed and treated on the day of sacrifice just like the animals sacrificed on PND 42 with regard to time of dosing, collection of vaginal opening and estrous cycle observations, etc. It is critical that sacrifices are completed by 1300 hours due to normal diurnal fluctuation in thyroid hormone levels. In order to minimize animal stress, the Guideline recommends sacrifices be performed in a separate room from the holding room and that the time from transfer between rooms to sacrifice be as brief as possible.

The preferred method of euthanasia is by injectable anesthetic, followed immediately by decapitation in order to obtain a sufficient volume of blood for the T_4 and TSH measurements. If necessary to obtain a sufficient volume of blood, the decapitation may be performed after anesthesia has been achieved but before death.

The main concern with use of anesthetic or asphyxiant is the induction of stress, which may affect hormone levels within a short period of time. Use of injectable anesthetic is preferred due to better delivery control and thus potentially shorter times to induce deep anesthesia than typically occurs with inhalational anesthetics.

A less preferred but still acceptable method of sacrifice is by decapitation without anesthesia. Decapitation has generally not been found to interfere with the integrity of the thyroid, which must be maintained in order to obtain thyroid weight and histology sections.

The Guideline recommends the order of necropsy be randomized or otherwise evenly distributed across all groups being necropsied that day. When two or more test chemicals use the same control group, it is particularly important to intersperse the control animal necropsies across the entire time span in which all of the necropsies for all the test chemicals and dose levels are conducted.

It is recommended that blood from the trunk of the animal be collected immediately into serum separation tubes, centrifuged and the serum stored at -20°C or colder for subsequent hormone and blood chemistry measurements.

The Test Guideline recommends that the following procedures be used during the necropsy. Organs to be collected include: the ovaries (without oviducts), uterus, thyroid (with attached portion of trachea), liver, kidneys, pituitary, and adrenals. Fresh wet weights are generally reported for each organ, except the thyroid/trachea and uterus. The thyroid is typically weighed after fixation. The kidneys, adrenals, and ovaries are generally weighed as pairs. The

uterus and cervix are generally separated from the vagina and weighed. The uterus is then typically placed on filter paper, slit to allow the fluid contents to leak out, gently blotted dry, and weighed again. Small tissues such as the adrenals and pituitary, as well as tissues that contain fluid, are typically weighed immediately to prevent tissues from drying out prior to weighing.

The Guideline recommends ovaries, uterus and kidneys be fixed in 10% buffered formalin for at least 24 hours, and stored in 70% ethanol until embedding in paraffin. They are then sectioned and stained with hematoxylin and eosin (H&E) for subsequent histological evaluations. The thyroid, with attached trachea, is fixed in 10% buffered formalin for at least 24 hours. The thyroid (with parathyroids) is then dissected from the trachea, blotted, weighed and stored in 70% ethanol.

2. Hormone Assays

Hormonal measurements can be conducted using radioimmunoassay (RIA), immunoradiometric assay (IRMA), enzyme-linked immunosorbent assay (ELISA), or time-resolved immunofluorescent procedures. The Guideline recommends that, regardless of which is used, multiple quality control (QC) samples run in duplicate be included among the test samples. The Guideline recommends that any measurement kit that is used be shown to yield appropriate values for control rats at the laboratory performing the pubertal assay. This includes demonstrating that QC was performed as described by the kit manufacturer and that the performance falls within the range defined by the manufacturer. If the kit does not provide or specify a standard control, then an additional option would be for the lab to use its own historical quality control samples. The DER should generally include the lab's criteria for evaluating the kit’s performance.

3. Blood Chemistry

Any standard panel of blood chemistry tests that includes creatinine and blood urea nitrogen would be scientifically appropriate as long as the measurements are calibrated for rats and the normal ranges for controls are reported. The normal ranges for controls may be from the literature (in which case the reference should be given in the DER), or from historical controls (which also should be included). Clinical chemistry parameters are typically performed on all animals, including controls, and are measured at terminal sacrifice.

Clinical chemistry levels are usually considered adverse when at least two liver parameters have a dose dependent, biologically significant change in albumin, alanine aminotransferase, aspartate aminotransferase, alkaline phosphatase, bilirubin, cholesterol or gamma glutamyltransferase. Typically, these changes should corroborate each other and be consistent with the known significance of the parameters. With renal toxicity, serum creatinine concentrations tend to parallel changes in blood urea nitrogen (BUN). Thus, in well-controlled toxicity studies in rodents, relatively small increases in serum BUN and creatinine concentrations (e.g., 1.5-fold) can be indicative of renal injury but significant and consistent increases in BUN or creatinine above control ranges, including laboratory reference ranges, provide more support for a treatment related effect.

4. Histology

The Guideline recommends uterus, thyroid, one ovary, and one kidney be evaluated for pathologic abnormalities and potential treatment-related effects. Thyroid sections are subjectively evaluated for follicular cell height and colloid area, preferably using a five point grading scale (1 = shortest/smallest; 5 = tallest/largest) and any abnormalities/lesions noted. The Guideline recommends that a minimum of two sections of each of the two lobes of the thyroid are evaluated. The Guideline provides example photomicrographs illustrating the magnitude of differences that are typically evaluated as separate scores. Ovarian histology includes evaluations of any abnormalities/lesions (such as ovarian atrophy) and follicular development, including: presence/absence of tertiary/antral follicles, presence/absence of corpora lutea, changes in corpus luteum development, changes in number of both primary and atretic follicles. The Guideline recommends that five random sections be evaluated using the method of Smith *et al*. (1991). This article was referenced to support the use of appropriately prepared random sections as opposed to serial sections. The follicle counts described by Smith *et al*. are not part of the pubertal assay. Other methodologies for obtaining a representative distribution of sections are acceptable but should be described in the DER. Uterine histology documents cases of uterine hyper- or hypotrophy as characterized by changes in uterine horn diameter and myometrial, stromal, or endometrial gland development. The histological assessment takes into account the stage of the estrous cycle of the female at the time of necropsy as ovarian and uterine cellular changes are dependent upon endocrine status.

I. Statistical Analyses

The reviewer should consider whether there are any data points that should be excluded from the data set, and whether any data points that are identified as statistical outliers should actually not be excluded. The decision to exclude data points should be based on statistical analysis and toxicological judgment. Values due to obvious technical errors should be excluded. The Guideline recommends the study report justify and report outliers in the raw data; the DER should explain the exclusion of any data points.

The Test Guideline recommends that the following procedures be used for statistical analyses. For additional guidance, please refer to the Corrections and Clarifications on Technical Aspects of the EDSP Tier 1 Assays (OCSPP Test Guideline Series 890) document. This document may be found by way of the EDSP web page (http://www.epa.gov/endo/). It is recommended that the data be analyzed for normal distribution and heterogeneity of variance. The Guideline recommends that all data except histology and cyclicity evaluations (i.e., initial body weight [PND 22], age and body weight at vaginal opening, body weight, body weight gain, and organ weights at necropsy, and serum hormones) be analyzed by Analysis of Variance (ANOVA).

For calculation of body weight gain, the Guideline recommends use of the body weight on the last day all the animals were weighed; specifically, if sacrifices were performed over two days, it would not be appropriate to use the day when only the last half of the animals were available. If the study was conducted in blocks, then the Guideline recommends that the analysis be performed using a two-way ANOVA with Block and Treatment as main effects.

Age and body weight at vaginal opening and all organ weights are also typically analyzed by Analysis of Covariance (ANCOVA), using the body weight at PND 22 as the covariate. When statistically significant effects are observed ($p < 0.05$), it is recommended that treatment means be examined further using the appropriate pairwise comparison tests that would be needed in order to identify dose groups that are significantly different from the control group. Where there is heterogeneity of variance, the recommended procedure is to transform data appropriately prior to ANOVA/ANCOVA, or to analyze the data using an appropriate non-parametric test. However, non-parametric analysis is the method of last resort as it does not allow analysis of covariance. In addition to ANOVA and ANCOVA, it is recommended that the unadjusted and adjusted values are typically examined for linear trend with dose level.

In cases where vaginal opening has not occurred prior to necropsy, the last day of observation +1 is generally used as the age at vaginal opening when determining the mean for each group. For example, if the animal was sacrificed on PND 42 without vaginal opening, PND 43 is used as the value for that animal when determining the mean for the treatment group.

Chi-square analysis is recommended to determine significant differences between the cycling status (cycling vs. not cycling) of the treated groups from the control group. Similarly, chi-square analysis is recommended to determine significant differences between treated groups from the control group for the percent of animals cycling regularly. Cycle length may be defined as either the number of days from one proestrus to the next proestrus, or from one diestrus to the next diestrus. Whichever definition is chosen, it should be applied uniformly to all groups in the study. Incomplete cycles should not be counted in calculating mean cycle length. Mean cycle length for each animal is calculated first, and the mean of these means is then calculated to represent the group.

IV. STUDY INTERPRETATION

The following sections and text summarize the information generally expected to be obtained by the study that would generally be relevant to evaluating the pubertal assay and therefore need to be included in each DER. The Guideline recommends the study report provide all the raw data and data summaries in electronic format (spreadsheet or comma-separated values), along with all formatting information that is necessary to read the data. In circumstances where the reviewer independently calculates summaries and statistics for endpoints, it is recommended that these electronic summaries be used to reduce errors. Use of the electronic raw data for these purposes, if provided/used should be documented in the DER. The DER should also have an executive summary describing the number and strain of rats used in the study, the dose levels and chemicals tested, and the effects, with levels of statistical significance for all endpoints.

A. Results

1. General Growth and Vaginal Opening

The DER should generally include the mean body weight ± standard deviation (SD) for each day during dosing for each treatment group, including vehicle control. The DER should generally provide the data in tables. In general, the p-value is reported in the DER when the difference is statistically significant. The numeric data for general growth and vaginal opening should typically also be provided in the DER in tabular form and include the mean, SD,

coefficient of variation (CV), number of animals (N), and p-value for the following endpoints, for each test-chemical dose group and control, both unadjusted (U) and adjusted (A) for body weight on PND 22.

- Age at attainment of vaginal opening
- Body weight at the age of attainment of vaginal opening
- Initial body weight (PND 22)
- Body weight on the last day all the animals were weighed
- Final body weight as percent of control (leave control column blank)
- Body weight gain from first dose to the (first) day of necropsy

Endpoints that show an effect (by ANOVA/ANCOVA or a non-parametric test, as appropriate) should be indicated in the DER. The DER should list any transformation used to eliminate heterogeneity of variance, or state if a non-parametric test was used. The pairwise test used to compare the means of dosed groups to the mean of controls should be indicated.

The incidence should also be provided in the DER of the proportion of animals in which vaginal opening had not occurred by the time of necropsy (e.g., X/15). For animals which have not attained VO at the time of necropsy, a value of necropsy day plus one (e.g., 43) will be used for those animals when calculating the group mean.

The summary table that should be used to report general growth data in the DER is shown below in Table 1. Dose group parameters that are significantly different from the vehicle control group ($p < 0.05$) will be indicated.

Table 1. General Growth and Vaginal Opening [a]

Parameter Evaluated		Vehicle Control				Low Dose ([#] mg/kg/day)				High Dose ([#] mg/kg/day)			
		Number of Animals Examined	Mean	SD	CV	Number of Animals Examined	Mean	SD	CV	Number of Animals Examined	Mean	SD	CV
Initial body weight (PND 22; g)	U												
	A												
Body weight at vaginal opening (g)	U												
	A												
Final body weight (g)	U												
	A												
Final body weight (% of control)	U												
	A												
Body weight gain (final – initial; g)	U												
	A												
Age at vaginal opening (PND)	U												
	A												
Proportion unopened (#/N)													

a Data were obtained from page [#] of the study report, and are the average of [#] female rats per dose group.
U = Unadjusted for body weight on PND 22
A = Adjusted for body weight on PND 22
SD = Standard Deviation
CV = Coefficient of Variation
* Significantly different from controls at $p<0.05$.
NA= Not applicable

2. Organ Weights at Necropsy

The DER should report the mean, SD, CV, number of animals (N), for both unadjusted (U) and adjusted (A; for body weight on PND 22) organ weights for the following organs: liver, kidneys, pituitary, adrenals, ovaries, uterus, and thyroid. The mean, SD, and p-value of the organ-weight-to-body-weight ratio for liver, kidney, adrenals, and pituitary is also provided. For ovaries, uterus, and thyroid weights, do not use relative organ to body weight ratios, and do not adjust for body weight at necropsy.

The summary table that should be used to report organ weight data in the DER is shown below in Table 2. Dose group parameters that are significantly different from the vehicle control group ($p < 0.05$) should be indicated.

Table 2. Organ Weights at Necropsy[a]

Organ		Vehicle Control				Low Dose ([#] mg/kg/day)				High Dose ([#] mg/kg/day)			
		Number of Animals Examined	Mean	SD	CV	Number of Animals Examined	Mean	SD	CV	Number of Animals Examined	Mean	SD	CV
Liver (g)	U												
	A												
	R												
Kidneys (g)	U												
	A												
	R												
Pituitary (mg)	U												
	A												
	R												
Adrenals (mg)	U												
	A												
	R												
Ovaries (mg)	U												
	A												
Uterus, wet (mg)	U												
	A												
Uterus, blotted (mg)	U												
	A												
Thyroid, fixed (mg)	U												
	A												

a Data were obtained from page [#] of the study report, and are the average of [#] female rats per dose group.
U = Unadjusted for body weight on PND 22
A = Adjusted for body weight on PND 22
SD = Standard Deviation
CV = Coefficient of Variation
R = Organ-to-body weight ratio (relative to body weight)
* Significantly different from controls at $p<0.05$.

3. Estrous Cyclicity

The DER should include detailed information on estrous cyclicity for all animals and summary information for the dose groups, including: (i) the mean age at first vaginal estrus, (ii) the mean cycle length for each group, (iii) the percent of each group cycling, (iv) the percent of each group cycling regularly, and (v) the stage of the cycle at the time of necropsy. For purposes of determining the cyclicity, an animal is considered to be "not cycling" if she shows three or more consecutive days of estrus or five or more consecutive days of diestrus. An example of a summary table that could be used to report estrous cyclicity data in the DER is shown in Table 3. As with all summary tables, dose group parameters that are significantly different from the vehicle control group ($p < 0.05$) should be marked.

TABLE 3. Estrous Cyclicity[a]

Dose Level (mg/kg/day)	Number of Animals (N)	Mean Age at First Vaginal Estrus (PND)	Mean Cycle Length (days)	Cycling (%)	Regularly Cycling (%)	Cycle Status at Necropsy (# Females)			
						Diestrus	Proestrus	Estrus	Not Cycling
Vehicle									
Low Dose									
High Dose									

a Data were obtained from page [#] of the study report, and are the average of [#] female rats per dose group.
* Significantly different from controls at $p<0.05$.

4. Thyroxine and Thyroid Stimulating Hormone Levels and Clinical Chemistry

The mean, SD, CV, number of animals, and p-value for the thyroxine (T_4) and thyroid stimulating hormone (TSH) levels, for each treatment group, including vehicle control should be reported. Similarly, the mean, SD, CV, number of animals and p-value (when statistically significant) for each of the clinical chemistry parameters measured should be presented in the DER. Also, the normal range for each parameter should be provided in the study report and appropriate data appended to the DER. It should be indicated whether these normal values are from the literature (provide reference) or from laboratory historical controls. A comparison will be made by the Reviewer for the normal values to those observed during the study.

The summary table that will be used to report hormone levels and clinical chemistry data in the DER is shown below in Table 4. Dose group parameters that are significantly different from the vehicle control group ($p < 0.05$) should be indicated.

TABLE 4. Hormone Levels and Clinical Chemistry [a]

Parameter Evaluated	Vehicle Control				Low Dose ([#] mg/kg/day)				High Dose ([#] mg/kg/day)			
	Number of Animals Examined	Mean	SD	CV	Number of Animals Examined	Mean	SD	CV	Number of Animals Examined	Mean	SD	CV
Hormones												
Serum T_4, Total (µg/dL)												
Serum TSH (ng/mL)												
Clinical Chemistry												
Creatinine (µmol/L)												
Blood urea nitrogen (mmol/L)												

a Data were obtained from page [#] of the study report, and are the average of [#] female rats per dose group.
SD = Standard Deviation
CV = Coefficient of Variation
* Significantly different from controls at $p<0.05$.

5. Histopathology

The uterus, thyroid, one ovary, and one kidney are evaluated for pathologic abnormalities and potential treatment-related effects. Thyroid sections should be subjectively evaluated for follicular cell height and colloid area, preferably using a five point grading scale (1 = shortest/smallest; 5 = tallest/largest) and any abnormalities/lesions noted. The Guideline recommends a minimum of two sections of each of the two lobes of the thyroid be examined in order to obtain representative sample of the thyroid tissue from each lobe.

The judgment of histopathological changes observed on the ovary, uterus, and/or thyroid glands are important to consider when evaluating the organ weights and hormone levels measured in this assay. Severity and incidences of effects and dose-response relationship may also be important information to consider. A summary of the incidence data for histological findings should be provided in the DER (as in Table 5), whereas the study report may additionally provide example photomicrographs of significant observations for the record.

Table 5. Incidence of Histopathological Lesions of the Thyroid Gland [a]

Treatment Groups	Parameter Evaluated											
	Colloid Quality			Follicular Cell Height (Increase)			Follicular Cell Height (Decrease)			Follicular Cell Shape		
	Severity	Incidence		Severity	Incidence		Severity	Incidence		Severity	Incidence	
		O	E		O	E		O	E		O	E
Vehicle Control	0			0			0			0		
	1			1			1			1		
	2			2			2			2		
	3			3			3			3		
Low Dose ([#] mg/kg/day)	0			0			0			0		
	1			1			1			1		
	2			2			2			2		
	3			3			3			3		
High Dose ([#] mg/kg/day)	0			0			0			0		
	1			1			1			1		
	2			2			2			2		
	3			3			3			3		

a Data were obtained from page [#] of the study report..
Thyroid histopathology is graded 0 – 3 based on severity: 0=Not remarkable, 1=Mild, 2=Moderate, 3=Severe.
See OECD No. 82 for reference.
O = Number Observed
E = Number Examined

Table 6. Incidence of Histopathological Lesions of the Uterus, Ovary and Kidney[a]

Findings	Dose Level (# mg/kg bw/day)					
	Vehicle Control		Low (#)		High (#)	
	# Observed	# Examined	# Observed	# Examined	# Observed	# Examined
Uterus						
[observation type]						
Ovary						
[observation type]						
Kidney						
[observation type]						

a Data were obtained from page ## of the study report..

B. Data Interpretation Procedure

The Guideline recommends that the highest dose level be at or just below the Maximum Tolerated Dose (MTD) but need not exceed the Limit Dose (1000 mg/kg/day); however, typically, the Agency also considers the toxicity profile of the chemical (i.e., cholinesterase inhibition, target organ toxicity, etc.) in dose selection. Typically the second dose level is spaced to produce a lesser degree of toxicity relative to the high dose unless justification is provided for testing at a different level.

Generally, negative results for interaction with the endocrine system in the pubertal assay will require demonstration that the highest dose level tested was at or near a dose that was previously determined to produce toxicity. To determine the acceptability of a study in which a positive finding occurs only at a dose level that causes clear adverse effects (e.g., $\geq$ 10% decrease in body weight gain at termination compared to controls), a weight-of-evidence assessment may be required.

As with the standard practice for evaluating changes in organ weights for the pituitary, liver and kidney, EPA recommends that terminal body weights be used to calculate relative organ weight changes. In contrast, because endocrine-active agents themselves may have an effect on body weight, it is typically appropriate to adjust for covariance with body weight (e.g. PND 22) before chemical treatment began.

Weight of ovaries and uterus are interpreted carefully, due to the natural variability in these endpoints in cycling animals. In general, regularity of cycling should be given more emphasis in a weight of evidence assessment than lack of statistical significance for the difference in weight of ovary or uterus in treated animals compared to controls. The judgment of the pathologist as to whether an effect on ovary, uterus, and/or thyroid is associated with exposure to the test chemical is important to consider when evaluating the organ weights and hormone levels measured in this assay. Severity and incidence of effect(s) can also be important information to consider.

Changes in hormones and histopathology may occur independent of changes in organ weights. The severity and incidence of effect(s) also are important considerations. Because there are multiple endpoints examined in this assay, there is redundancy for the detection of potential endocrine system interaction. Redundancy is particularly useful when the responses from all the similar endpoints are consistently positive, as it gives greater confidence that the interaction with the endocrine system is real. However, consistency across all redundant endpoints is not required in order to infer interaction with the endocrine system. There may be valid reasons for apparently-redundant endpoints to differ in their response.

As with other standard evaluation procedures for OCSPP Test Guideline studies, the concurrent control is generally the first order comparison for the treatment groups. The laboratory may also choose to provide appropriate historical control data based on equivalent study parameters (strain, age, endpoints). The adequacy of the concurrent control data may be evaluated by comparison to the historical control data and/or the performance criteria provided in the Test Guideline. The data for the concurrent controls should generally fall within the range of specified performance criteria (see Table 7 below). The performance criteria are indications of whether the sensitivities of individual endpoints are sufficient to allow conclusions that the test chemical did not affect those endpoints.

Additional guidance on data interpretation are provided in the OCSPP Guideline 890.1450 for the Female Pubertal Assay and in the Agency's document titled "Corrections and Clarifications on Technical Aspects of the EDSP Tier 1 Assays (OCSPP Test Guideline Series 890)" which can be found on the EDSP web page (http://www.epa.gov/endo/).

C. Performance Criteria

The following performance criteria have been established for the vehicle-control animals. The objective of the performance criteria is to ensure the sensitivity of the endpoints relative to other confounding factors during the course of the study (e.g., phytoestrogens in diet, contaminants in drinking water). See the Data Interpretation Procedure above for correct use of the performance criteria. Units for the endpoints are shown in Table 7. The "Mean," "2 SDs," "CV," and "1.5 CV" columns describe the mean, two standard deviations, coefficient of variation, and 1.5 times the coefficient of variation for the given endpoints in the historical controls. The targeted mean values and CVs for the vehicle control group for various endpoints are provided below.

Table 7 below is intended to be used as a worksheet for comparison of the observed results to the performance criteria. Report the relevant findings in the DER.

Table 7. Performance Criteria for Controls (Sprague-Dawley strain)

Endpoint	Mean	2 SDs	Acceptable Range	CV (%)	1.5 CV (%)	Top of Acceptable CV Range[a]
Uterus, blotted (mg)						
	298.89	111.49	187.40 to 410.38	28.26	9.48	37.73
Ovaries (mg)						
	75.65	39.11	36.54 to 114.77	17.37	5.83	23.20
T_4 (total; µg/L)						
	4.03	1.34	2.69 to 5.38	21.38	8.01	29.39
Thyroid weight (mg)						
	14.20	8.00	6.20 to 22.20	23.97	14.61	38.58
Age at vaginal opening (PND, where day of birth = PND 0)						
	33.15	2.47	30.67 to 35.62	4.69	1.82	6.52
Weight at vaginal opening (g)						
	116.58	14.86	101.71 to 131.44	9.81	4.16	13.97
Final body weight (g)						
	154.70	49.85	104.86 to 204.55	6.69	2.25	8.93
Adrenals (mg)						
	43.59	5.25	38.34 to 48.84	16.02	6.95	22.97
Kidneys (g)						
	1.57	0.63	0.95 to 2.20	8.42	2.34	10.76
Liver (g)						
	8.05	3.73	4.32 to 11.78	9.58	3.55	13.13
Pituitary (mg)						
	8.97	3.11	5.86 to 12.08	15.76	11.22	26.97

a Bottom of the acceptable range for coefficient of variation is zero.
SD = Standard Deviation
CV = Coefficient of Variation

Note: No performance criteria have been established yet for TSH as there were too few studies from which reliable historical control values resulting from the same analytical method could be obtained. Such criteria may be established in the future as more data become available.

V. CHARACTERIZATION OF FINDINGS

On completion of the review of this assay, the Agency will conduct a weight of evidence analysis to consider the potential of the chemical to disrupt the estrogen, androgen, or thyroid hormone systems. Chemicals with demonstrated evidence of a potential to interact with the estrogen, androgen, and/or thyroid hormone systems will be considered as candidates for Tier 2 testing.

VI. DATA EVALUATION RECORD

Once the study has been reviewed using the principles described in the previous sections of this SEP, a DER will be prepared. A DER template is available that provides additional guidance for the preparation of the DER.

VII. REFERENCES

Cooper R.L., *et al*. (1993) Monitoring of estrous cycle in the laboratory rodent by vaginal lavage. In: Methods in Toxicology. Vol. III, Part B. Female Reproductive Toxicology. Edited by Chapin RE and Heindel J. Academic Press: Orlando. p. 45-56.

Goldman, J.M., *et al*. (2007) The rodent estrous cycle: characterization of vaginal cytology and its utility in toxicological studies. *Birth Defects Research (Part B)* 80: 84-97.

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/)

Markaverich BM, *et al*. (2002) Identification of an endocrine disrupting agent from corn with mitogenic activity. *Biochem. Biophys. Res. Commun.* 291(3): 692-700.

OECD (2007) Guidance Document on Amphibian Thyroid Histology. Environmental Health and Safety Publications. Series on Testing and Assessment. No. 82. Paris

Smith B.J., *et al*. (1991) Comparison of random and serial sections in assessment of ovarian toxicity. *Reproductive Toxicology* 5(4): 379-383.

USEPA, Office of Prevention, Pesticides, and Toxic Substances. August 1998. Health Effect Test Guidelines 870.3800: Reproduction and fertility effects. EPA 712-C-98-208.