Chapter Five

Screening and Testing

Table of Contents

Figures

Tables

Appendices

I. Chapter Overview

This chapter describes the EDSTAC recommendations regarding development of a screening and testing program for assessing the potential of pesticides and other chemicals to disrupt endocrine function in humans and wildlife. Where appropriate, strengths and limitations of options are discussed and possible future research projects are identified to develop needed procedures. The EDSTAC established the Screening and Testing Work Group (STWG) [see Appendix D for a list of work group members] to assist in their efforts to provide guidance to EPA regarding the development and implementation of its endocrine disruptor screening and testing strategy. The STWG work formed the basis for this chapter and recommendations. References, for all sections of the chapter, and additional sources can be found in Appendix J.

After this introduction, the chapter is comprised of seven main sections: first, the concepts and design parameters involved with Tier 1 Screening (T1S); second, the Proposed T1S Battery; third, the general principles in evaluating tier 1 and tier 2 results; fourth, the concepts and design parameters involved with Tier 2 Testing (T2T); fifth, the Proposed T2T Battery; sixth, a summary of the interconnections between components of the EDSTP; and finally, a discussion of Standardization, Validation, Methods Development, and Research.

The T1S sections begin with an explanation of the purpose of screening and identification of five criteria used to design the screening battery. An outline of the proposed T1S battery follows and overviews of each recommended assay are then provided, along with discussions of the value of including both *in vitro* and *in vivo* assays. Four alternative assays for consideration are also discussed. Finally, a section on evaluating the battery includes a discussion of a weight of evidence approach to evaluating T1S.

In developing the T1S battery, the EDSTAC considered endpoints for their utility in screening CSMs for their potential to interact with the endocrine systemdisrupting CSMs. The goal of T1S is to detect the potential for chemical substances or mixtures (CSM) to affect estrogen, androgen, or thyroid hormone activities. Assessing these activities is relevant as changes in them may adversely affect the development, reproductive function, or chronic health status of people or animals. The objective of T1S is not to determine dose-response relationships, confirm the mechanism of action, or determine the adversity of the chemical's effect on reproduction and/or development; however, screening assays must be sensitive enough to detect all known xenobiotics that act via the mechanism of action each assay is designed to detect.

The screening battery presented here has been designed to ensure effects on the aforementioned hormonal modalities will be detected. There are instances in which decisions had to be made as to whether to select an assay that was highly specific for a hormonal activity versus one that may be less specific but more sensitive and apical (i.e., a more comprehensive assessment of functions that are relevant to reproduction, development, or chronic health). In those instances we opted for the latter, as this strategy: 1) better fulfills the first criterion for screens (that they be sensitive); and 2) is better aligned with the overall mission of detecting health hazards regardless of mechanism of action. Each recommended

assay is described later in the chapter. These assays require varying levels of additional development, standardization, and validation.

The T2T sections begin with discussions of the purpose of testing, guidance for selecting Tier 2 tests, and the issue of dose considerations for components of T2T. An outline of the proposed T2T battery is followed by overviews of the mammalian two-generation reproductive toxicity study, alternative less comprehensive mammalian tests, and tests using other vertebrate and invertebrate taxa. A summary of the interconnections between HTPS, bypassing T1S, low dose concerns, and the definitiveness of T2T follows. Finally, recommendations regarding implementation of the standardization, validation, and research program are presented.

The goal of T2T is to determine whether a substance is an endocrine disruptor for EAT verify that observations made in T1S occur and to determine the adverse consequences to the organism of the activities observed in T1S and the dose response relationships. This is done in the larger context of testing for reproductive and developmental toxicity potential by any mechanism (including EAT) using apical study designs.

II. Tier 1 Screening Concepts and Design Parameters

As discussed in Chapter Two, CSMs can alter endocrine function by affecting the availability of a hormone to the target tissue, and/or affecting the cellular response to the hormone. Mechanisms regulating hormone availability to a responsive cell are complex and include hormone synthesis, serum binding, metabolism, cellular uptake (e.g., thyroid), and neuroendocrine control of the overall function of an endocrine axis. Mechanisms regulating cellular response to hormones are likewise complex and are tissue specific. Because the role of receptors is crucial to cellular responsiveness, specific nuclear receptor binding assays are included. In addition, tissue responses that are particularly sensitive and specific to a hormone are included as endpoints for Tier 1 Screens.

For purposes of clarity the following definitions are utilized in this chapter. Estrogenic refers to compounds whose effects are mediated through the estrogen receptor (ER), initiating a cascade of cell/tissue specific effects similar to those initiated by estradiol, as well as those resembling estrogen but are not or have not been shown to be mediated through the ER. Similarly, androgenic effects are androgen receptor (AR) mediated, as opposed to androgen-like effects, which may not be mediated via AR. In contrast, the terms antiandrogenic and antiestrogenic are not specifically limited to AR- and ER-mediated interactions. In this context, agonists bind to the receptor and act like the endogenous hormone; antagonists bind to the receptor and act opposite to the endogenous hormone. Antihormones can act via: 1) the steroid hormone receptor; 2) steroid hormone synthesis inhibition; 3) reduction of bioavailability by reducing the amount of free hormone in the serum; 4) increased hormone metabolism leading to reduced serum hormone levels; and 5) other mechanisms.

A. Introduction to T1S

As mentioned in Chapter Four, the number of chemicals needing evaluation is huge. T1S is intended to make the evaluation process more efficient by distinguishing those chemical substances and mixtures which have potential endocrine activity from those that have little or no such potential. The EDSTAC considered all known endocrine disruptors in developing the T1S battery and believe that the proposed battery, if validated, will have the necessary breadth, and depth to detect all known disruptors of EAT.Consensus was reached that the proposed T1S battery, when validated, will detect all known disruptors of estrogen, androgen, and thyroid. Therefore, following application of the T1S battery, a chemical substance or mixture (CSM) will be designated as having either: 1) the potential for estrogen, androgen, or thyroid activity, which will require further analysis in T2T to verify and evaluate that potential or 2) low or no potential for estrogen, androgen, or thyroid activity, which will allow assignment of CSMs to the "hold unless ..." box (see Section IV. in this chapter, on general principals in evaluating tier 1 and tier 2 results for further discussion of how this decision is made).

In developing the proposed T1S battery, many existing and potential assays were evaluated for their relative strengths and weaknesses (one-page overviews of evaluated assays can be found in Appendix K). The proposed T1S battery contains mammalian *in vitro* and *in vivo* assays and *in vivo* nonmammalian assays. T1S is designed to be sufficiently sensitive to identify chemicals which may be endocrine disruptors in humans and wildlife. In addition, T1S results should inform T2T, in terms of providing guidance on which tests to perform, which endpoints to include, which organisms to assess to satisfy the goals of T2T, and to assist in determining the range of doses to be used. These goals include identifying those doses (dose-response), life stages (most sensitive), and organisms (most appropriate, sensitive, and at risk) in which adverse effects are likely to occur.

B. Criteria for T1S

The T1S battery proposed by EDSTAC has been developed such that, at the completion of the selected assays, the EPA and other stakeholders will accept, both scientifically and as a matter of policy, the assignment of CSMs as either having: 1) low or no potential for estrogen, androgen, or thyroid endocrine disruption, or 2) as having such potential. The ability to accept either outcome requires that the chosen T1S battery meets the five criteria identified below.

1. The T1S battery should maximize sensitivity to minimize false negatives while permitting an as of yet undetermined, but acceptable, level of false positives. This criterion expresses the need to "cast the screening net widely" in order not to miss potential endocrine disruptors.

2. The T1S battery should include a range of organisms representing known or

anticipated differences in metabolic activity. The battery should include assays from representative vertebrate classes to reduce the likelihood that important pathways for metabolic activation or detoxification of parent CSMs are not overlooked.

3. The T1S battery should be designed to detect all known modes of action for the endocrine endpoints of concern. All chemicals known to affect the action of estrogen, androgen, or thyroid hormones should be detected.

4. The T1S battery should include a sufficient range of taxonomic groups among the test organisms. There are known differences in endogenous ligands, receptors, and response elements among taxa that may affect endocrine activity of CSMs.

5. The T1S battery should incorporate sufficient diversity and complementary among the endpoints and assays to reach conclusions based on "weight-of-evidence" considerations. Decisions based on the battery results will require weighing the data from several assays.

The T1S must be relatively fast and efficient while meeting the criteria described above. The EDSTAC recommends that if changes are made to the proposed T1S battery, based upon development of new, validated assays, the "amended" battery still needs to meet these criteria.

III. Proposed Tier 1 Screening Battery

A. Outline of Proposed T1S Battery and Possible Alternatives

1. Proposed T1S Battery

The T1S battery proposed by the EDSTAC includes the following three *in vitro* assays, three *in vivo* mammalian assays, and two *in vivo* nonmammalian assays. Based on existing data, the EDSTAC believes this battery will detect EAT activity, provided all of the component assays can be properly developed, standardized, and validated.

in vitro

1. ER Binding/ or Transcriptional Activation Assay;
2. AR Binding/ or Transcriptional Activation Assay; and
3. Steroidogenesis Assay with Minced Testis.

in vivo

1. Rodent 3-Day Uterotrophic Assay (Subcutaneous);
2. Rodent 20 Day Pubertal Female Assay with Thyroid;
3. Rodent 5-7 Day Hershberger Assay;

4. Frog Metamorphosis Assay; and
5. Fish Gonadal Recrudescence Assay.

2. Alternative Assays for Possible Inclusion

In addition, the EDSTAC has identified an *in vitro* assay and three *in vivo* assays as possible substitutes, if properly developed, standardized, and validated, for some of the component assays in the proposed battery. These assays are:

in vitro

1. Placental Aromatase Assay.

in vivo

1. Modified Rodent 3-Day Uterotrophic Assay (Intraperitoneal);
2. Rodent 14-Day Intact Adult Male Assay With Thyroid; and
3. Rodent 20-Day Thyroid/Pubertal Male Assay.

Combinations of the alternative assays, if validated and found to be at least as sensitive for all endpoints as the assays in the proposed T1S battery, could potentially replace three of the component assays in the proposed T1S battery (*in vitro* steroidogenesis assay with testis, 20-day pubertal female assay, and 5-7-day Hershberger assay), thereby possibly reducing the overall time, cost, and complexity while maintaining equivalent performance of the overall T1S battery.

One alternative battery would include the ER binding or transcriptional activation assay, the AR binding or transcriptional activation assay, the modified rodent 3-day uterotrophic assay (ip), the rodent 14-day intact adult male assay with thyroid, the frog metamorphosis assay, the fish gonadal recrudescence assay, and, possibly, the placental aromatase assay.

The other alternative battery would include the ER binding or transcriptional activation assay, the AR binding or transcriptional activation assay, placental aromatase assay, the rodent 3-day uterotrophic assay (sc), the rodent 20-day thyroid/pubertal male assay, the frog metamorphosis assay, and the fish gonadal recrudescence assay.

Table 5.1 identifies the assays included in each of the batteries, the proposed and the two alternatives. In addition, Table 5.2 shows the assays in relation to which of the biological activities that may be affected by exogenous agents and lead to E, A, or T-related toxicity they are expected to detect.

Table 5.1

Assays Included in Proposed T1S Battery and Possible Alternatives

Assays	Proposed T1S Battery	Possible Alternative (1)	Possible Alternative (2)
in vitro			
Estrogen receptor binding	x	x	x
Androgen receptor binding	x	x	x
Steroidogenesis	x		
Placental aromatase		?	x
in vivo			
3-day uterotrophic	x (sc)	x (ip)	x
20-day pubertal female	x		
Hershberger	x		
14-day intact male		x	
20-day pubertal male			x
Frog metamorphosis	x	x	x
Fish gonadal recrudescence	x	x	x

Table 5.2

T1S Assays Related to Biological Activities Detected

		Anticipated to Detect:					
Assay	in Option	Estrogen Agonism	Estrogen Anagonism	Androgen Agonism	Androgen Antagonism	Thyroid-Related Effects	Ster Synth
in vitro							
Estrogen receptor binding	1,2,3	X	X				
Androgen receptor binding	1,2,3			X	X		
Steroidogenesis	1						X
Placental aromatase	3, 2?						
in vivo							
3-day uterotrophic	1,2,3	X		$(X)_2$			
20 day pubertal female	1	X	X			X	X
Hershberger	1			X			
Hershberger + T	1	$(X)_4$			X		
14-day intact male	2	X			X	X	X
20-day pubertal male	3	X		X	X	X	X

Frog metamorphosis	1,2,3	X	?[6]	?[6]	?[6]	X	X
Fish gonadal recrudescence	1,2,3	X	X	X	X	?[6]	X

Notes:
[1] HPG – indicates that the model has an intact hypothalamic-pituitary-gonadal axis (except Hershberger), and that effects on hypothalamic-pituitary control of gonadal endocrine function would be evaluated.
[2] It is likely that aromatizable androgens would be detected in this assay; however, given that there are no examples of environmental androgens, this point cannot be empirically demonstrated.
[3] Agents that affect LH level would be detected in the assay.
[4] Empirical demonstration that the assay detects estrogens is limited. The biology of the system suggests that they will be detected.
[5] Empirical demonstration that aromatase inhibitors are detected is limited. If sensitivity to aromatase inhibitors is lacking, a placental aromatase assay would be added to this option.
[6] The biology of these organisms suggests that these effects may be detectable. However, there are no empirical data to support the sensitivity of the assay for these endpoints.

3. Validation of the Battery

In order to provide sufficient data to allow informed decisions about the relative merits of the proposed T1S battery component assays and alternative assays (based on sensitivity, specificity, technical complexity, inter- and intra-laboratory variability, time, and cost), EDSTAC recommends that validation studies be initiated on all of the assays in the proposed battery as well as the alternatives.

If the assays comprising either of these alternative combinations are validated, EDSTAC recommends performance of the alternative battery containing these assays using the same standard test substances recommended for validation of the initial battery, which were selected on the basis of predetermined criteria (see Section VIII. F., Instituting a Validation Program). Sufficient information could then be available to allow an informed choice between the recommended battery or a variation, including the alternative assays, as the preferred T1S battery. This approach would be most expedient in the event that one or more of the recommended battery assays cannot be properly standardized and validated, since information would be immediately available on the alternative assays. EDSTAC believes this process provides a model for validation and incorporation of new assays, as they may be developed and proposed, into the T1S battery.

EDSTAC believes it is critical to acknowledge that the state-of-the-science, with respect to assay development and species selection, in this area is rapidly evolving, and bioassays are currently being developed that may offer distinct advantages over those assays and species presently some of those proposed for use. This is particularly the case for selection of non-mammalian species currently recommended for use in *in vivo* assays. Specific bioassays and species should be selected on a performance-based approach. As improve bioassays and/or more appropriate species are developed and validated, EDSTAC strongly encourages their use as assays for screening and/or testing. As they are developed and validated, the EDSTAC strongly encourages the use of these new or improved assays for screening. Select Some of the assays identified as research priorities by the EDSTAC are discussed in Section VIII. E. of this chapter.

Given the wide range of mammalian and avian species that may be adversely affected by endocrine disruptors, continued development of screens and tests is particularly important to ensure that a representative range of species and potential endocrine-related effects can be evaluated.

4. Assays Not Included in T1S

At this point in time, there are no data available to suggest that thyroid effects of CSMs are mediated through the receptor. Therefore, the proposed T1S battery does not currently include a thyroid receptor (TR) binding and/or transcriptional activation assay. Nevertheless, the EDSTAC is recommending that the HTPS program include evaluation of the TR. We believe including the thyroid assays in the HTPS program will enable EPA, and others, to obtain a better understanding, at relatively low costs, of whether effects could be mediated through the TR.

As mentioned earlier, brief overviews of all assays considered by the STWG can be found in Appendix K. In addition, Appendix M includes more thorough discussions of assays that the work group considered in detail, but decided not to include in the proposed T1S battery. EDSTAC's thoughts on the role of invertebrates in T1S can be found in Appendix N.

5. *In Utero* or *In Ovo* Exposure

One major EDSTAC concern was that the screening and testing program adequately assess the potential of a CSM to affect normal growth and development including reproductive structures and functions of offspring after embryonic *in utero* or *in ovo* exposures. An adequate evaluation must involve prenatal or prehatch exposure and retention of offspring through puberty to adulthood and reproductive performance. This evaluation, by necessity, is neither short term nor inexpensive. Anything less will not adequately assess the potential risk and may result in false negatives. There are also no known endocrine disruptors or reproductive toxicants that affect the prenatal conceptus in the absence of any effects on the adult, although the effective doses and affected endpoints may differ between the two life stages. Because of these two factors, the necessity to perform an adequate evaluation and the absence of any indication that use of an adult animal (at appropriate doses) will miss a potential endocrine disrupting CSM, the EDSTAC decided not to include any full life cycle studies (with embryonic *in utero* or *in ovo* exposure and evaluation of the adult offspring) in the proposed T1S battery (a possible protocol of an *in utero* exposure assay can be found in Section VII. E. of this chapter). However, full life cycle assessments are included in the proposed T2T battery for mammals, other vertebrates, and for invertebrates. These tests will employ a full range of doses, embryonic *in utero/in ovo* exposures, rearing offspring to adulthood, and a full complement of reproductive and developmental endpoints.

6. Methods to Select the Single Dose for *In Vivo* Assays

All T1S *in vitro* assays (including the steroidogenesis assay) will involve multiple doses, whether performed by HTPS or bench level methods, so a dose-response curve and assessment of relative potencies can be developed. Results from the HTPS (or its equivalent) will provide potency information (i.e., EC 50) relative to a positive control such as 17 beta estradiol (E2), diethylstilbestrol (DES), testosterone, or T4 for those CSMs which bind to the E, A, or T receptor. Information on the *in vivo* effective doses of E2, DES, testosterone, or T4 can be used to set a single high dose for the remaining T1S assays for these CSMs. It must be noted that there are no current data which indicate that thyroid toxicants act via binding to the thyroid hormone receptor(s). Thus, the proposed *in vitro* receptor binding/transcriptional activation assay may not inform dose selection for *in vivo* T1S assays for thyroid endpoints. For these CSMs, prior information and range-finding studies will be critical.

The EDSTAC recommends using one dose in the performance of the *in vivo* assays. Information to assist in selecting this single dose includes:

prior information, such as that available during the priority setting phase;
results from the HTPS (or its equivalent bench-level assays); and
results from range-finding studies, specifically performed for T1S.

Alternatively (especially for CSMs which are negative in the HTPS), a range-finding study can be performed at multiple doses (at least five) with a few animals per dose and a limited number of relevant endpoints. Range-finding studies specifically performed for each *in vivo* TIS assay will include the following:

use of the same species strain, sex(es), and age as in the T1S assay;
use of the same route of administration, vehicle, and duration of dosing as in the T1S assay;
use of multiple doses; the number of doses will depend on the availability and extent of prior information;
use of multiple animals per dose which may be fewer than the number used per group in the T1S assay;
use of relevant endpoints, which may be more limited than those in the T1S assay. For example, the range-finding study for the T1S uterotrophic assay may employ only body weights and uterine wet weight, while the T1S assay may also evaluate uterine gland height, serum hormone levels, and/or vaginal cornification, etc.
use of comparable animals, e.g., ovarectomized females for the uterotrophic range-finding study or castrated males for the Hershberger range-finding assay. However, there may be circumstances under which exceptions occur, e.g., use of intact males in the range-finding study for the Hershberger assay to define doses producing systemic toxicity and any effects on the reproductive system as a first pass approximation.
use of more than one range-finding study if the initial version does not identify the one dose to be used in the specific T1S assay if necessary by extrapolation or interpolation.

The dose to be selected for the *in vivo* assays should not result in excessive systemic toxicity, but should result in effects useful for detection of potential EAT disruption. However, in no case should the dose used be higher than one gram/kilogram body weight/day (i.e., the "limit" dose). The rationale for selection of doses for each range-finding study, all of the results for such studies, and the logic employed to select the single dose for the T1S assay should be included in the submission of T1S results for evaluation by the Agency as to the appropriateness of the study design, conduct, and conclusions. The results of range-finding studies' should be included in the submission of T1S results for evaluation as to the studies appropriateness by the Agency.

7. Routes of Administration

The route of administration for the proposed uterotrophic assay is subcutaneous (sc) injection while the route for the modified uterotrophic assay and 14-day intact adult male assay with thyroid is intraperitoneal (ip) injection. The route for all other mammalian *in vivo* assays is gavage (orogastric intubation). The parenteral (non-oral) routes avoid the first-pass metabolic effect of the liver and will permit detection of potential EDCs that are active as parent compounds and which undergo significant first-pass metabolism. Hepatic xenobiotic metabolism does occur eventually after parenteral administration (substantially with ip), so the potential effects of metabolites will be evaluated as well by these routes. Compounds are occasionally metabolized by the gut microflora; this type of metabolism has been shown to be important for some plant-derived estrogens. The oral route of exposure will allow for this type of metabolism.

B. Scientific Basis for *In Vitro* Screening for Estrogen, Androgen, and Thyroid Activities

General agreement has been reached on the strengths and limitations of most currently available *in vitro, in vivo,* and *ex vivo* methods for detection of toxicants that act via ER, AR, steroid hormone synthesis inhibition, and/or altered hypothalamic-pituitary-gonadal (HPG) mechanisms. With this in mind, several short-term *in vitro* assays for AR, ER, and steroidogenesis inhibition (SI) action were identified as quite useful in screening. *In vitro* methods also include steroidogenic enzyme/hormone synthesis, biochemical assays, and *in vitro* and testis steroid hormone synthesis.

Advantages of *In Vitro* assays include*:*

sensitivity to low concentrations increases detectability;
high specificity of response;
low cost;
small amount of CSM required;
in vitro assays can be automated, including use of robotics;

high throughput assays (thousands/month) can be developed;
results can be coupled with QSAR models and for database screening;
can be used for complex mixtures (sludge, water contaminants); and,
reduces or replaces animal use.

EDSTAC recognizes two categories of *in vitro* assays that may be used in T1S to assess the binding of test substances to receptors, i.e., cell-free assays for receptor binding and transfected cells designed to detect transcriptional activation. The EDSTAC believes receptor binding and/or functional assays should be included in T1S. The specific assays chosen, whether done "at the bench" or through the high throughput pre-screening process (discussed in detail in Chapter Four, Sections II. H., and V., should have the following characteristics:

evaluate binding to estrogen, androgen, and thyroid nuclear receptors;
evaluate binding to the receptor in the presence and absence of metabolic capability (e.g., one or more of the P450 isozymes, CYP1A1, CYP3A4, etc.);
distinguish between agonist and antagonist in functional assays; and
yield dose responses for relative potency of CSMs with endocrine activity.

Receptor binding assays should be performed for estrogen, androgen, and thyroid receptors if high throughput procedures are used, while if the assays are done at the bench level, only estrogen and androgen receptor assays are recommended and/or functional assays should be performed for estrogen, androgen, and thyroid receptors (specifically recommended is a stably transfected cell line like the MVLN cell line, if available, to assess transcriptional activation). If stably transfected cell lines are not available, then transiently transfected reporter gene assays should be used. MCF-7 proliferation assays are also acceptable; however, yeast-based assays are not recommended at this time. These assays can be performed either high throughput or at the bench level.

Receptor binding assays can use rat, mouse, or human ER or AR. These assays evaluate the ability of the xenobiotic CSM to displace the radio labeled endogenous ligand from the binding site, in a cell-free or whole cell system. Relative potency can be determined for positive CSMs. Assay limitations are solubility in the culture medium, inability to distinguish agonists from antagonists, lack of metabolic capability, and risk of degradation of the receptor.

The functional assay, specifically transcriptional activation, requires, for agonist or antagonist activity, that the CSM bind to the receptor. In addition, there is a consequence to the binding, i.e., transcription (synthesis of mRNA) of a reporter gene and translation of the mRNA to an identifiable detectable protein such as firefly luciferase or beta-galactosidase. In the case of the firefly luciferase, with substrate and cofactors present in the culture, there is a light flash detected from formation of the product when the enzyme is synthesized in response to transcriptional activation and acts on the provided substrate. In the case of the beta-galactosidase, with substrate and cofactors present in culture, the product is detected colorimetrically when the enzyme is synthesized in response to transcriptional activation and

acts on the provided substrate. The assay uses intact cells and may use different cell lines for assessment of effects on EAT binding domains with transfected (transiently or permanently) receptors and reporter gene constructs. This assay can distinguish between agonists and antagonists. Assay limitations are solubility, toxicity, permeability of the cell membrane, and lack of or limited metabolic capability. If a CSM must be metabolized to an active moiety, it will not be detected unless the limited residual metabolic capacity of the cultured cells is sufficient to transform the chemical to its active form. Metabolic activity might be provided by either preincubating the CSM with an S9 fraction (supernatant from 9000g x centrifugation of homogenized liver from a metabolically induced rat) or incorporating the S9 fraction into the treatment mixture. In addition, cell lines are being genetically engineered to incorporate genes for P450 enzymes as a method for extending their metabolic capacity and, perhaps, obviate the need for use of the S9 fraction.

For assessing receptor binding *in vitro*, EDSTAC is recommending both the receptor binding assays and the transcriptional activation assays be incorporated into the T1S battery, and subjected to validation and standardization. There is agreement that the transcriptional activation assays can provide more information than the receptor binding assays, since they measure not just binding capacity but also the physiological and biochemical consequences of that binding;. However, the limited database on the relative utilities of receptor binding and transcriptional activation assays do not allow EDSTAC to recommend one category of assay over the other at this time. however, the EDSTAC is not certain all of the transcriptional activation assays can be executed in all laboratories, as it is relatively new technology. Including the receptor binding and transcriptional activation assays in the standardization and validation program is expected to provide the data needed to reach a decision on whether both assays should be required or, if not, whether the receptor binding or transcriptional activation is preferred. should give EPA an answer to this question. If the transcriptional activation assays become standardized and validated, performance of the receptor binding assays will no longer be necessary. It is important to keep in mind that these assays evaluate just one of the possible mechanisms of endocrine disruption; if a CSM acts via another mechanism than the receptor, it will not be detected in these assays.

Large-scale high throughput pre-screening (HTPS) programs for CSMs have been employed, using standardized *in vitro* functional assays (i.e., transcriptional activation of a reporter gene), in the pharmaceutical industry. Several companies involved in drug design routinely screen chemicals for hormonal activity on a large scale (thousands per month).

In vitro evaluations, it is acknowledged, can provide both false positive and false negative results. False positives (i.e., active *in vitro* but not *in vivo*) arise *in vitro* when a chemical is not absorbed or distributed to the target tissue, is rapidly metabolically inactivated and excreted, and/or when some other form of toxicity predominates *in vivo*. False negatives are considered to be of greater concern if *in vitro* tests were used to the exclusion of *in vivo* methods. *In vitro* evaluations can result in false negatives due to their inability, or unknown capacity, to metabolically activate toxicants. As a result, the EDSTAC's recommended battery includes i*n vivo* methods in conjunction with *in vitro* techniques. Nevertheless, some *in vitro* assays may offer distinct advantages over *in vivo* assays when investigating the

activity of specific metabolites.

C. *In Vitro* Assay Overviews

The EDSTAC is proposing a specific assay for each of the ER receptor binding, ER transcriptional activation, AR receptor binding, AR transcriptional activation, and steroidogenesis categories in order for standardization and validation to occur efficiently. The receptor binding and transcriptional activation assays would be performed only on those CSM's not going through HTPS, while the steroidogenesis assay would be performed on all CSM's going through T1S. Equivalent assays could replace these if they meet specific performance criteria and were similarly validated. Even if HTPS is implemented, standardization and validation of these additional *in vitro* assays would allow them to be conducted in individual labs on a more limited basis. The following assays are the specific ones recommended for inclusion in the standardization and validation program. The following assays are the specific ones recommended for inclusion in the standardization and validation program.

1. Estrogen Receptor Assays
   ER Binding: Cell-Free ER Alpha Binding
   ER Transcriptional Activation: MVLN

2. AR Assays
   AR Binding: Cell-Free AR Binding
   AR Transcriptional Activation: AR Transcriptional Activation

3. Steroidogenesis
   Minced testis

1. Estrogen Receptor Assays

In vitro rat ER binding assays provide a rapid and fairly inexpensive method for quantifying the ability of chemicals to compete with DES or estradiol for ER. The assay can be used for measuring ER in cell-free extracts obtained from various tissue homogenates following *in vivo* exposure to an environmental chemical. In addition, the assay may be used to determine the ability of a given compound to compete with radio-labeled estradiol for binding to the ER. The technical aspects of the ER binding assay are well documented for receptors obtained from cytosolic or nuclear extracts of various mammalian and other vertebrate tissues (Anderson et al., l972; Korach et al., l979). In brief, cytosolic or nuclear extracts containing ER are incubated with [^{3}H] estradiol for 18 hours at 4^o C in the presence or absence of increasing concentrations of radio-inert DES or test chemicals. Nonspecific binding is assessed by the addition of 100 molar excesses of radio-inert DES. Bound [^{3}H]-and free

ligands are separated using hydroxyapatite extraction, or charcoal-dextran adsorption, and are quantified by scintillation counting.

The ER binding assays are less sensitive than the functional assays (IC50=), short-term duration, and can be standardized between laboratories. The assay is useful for evaluating effects of a test compound on ER distribution and number following *in vivo* exposure. In addition, the assay can be used to rapidly evaluate test compounds for their ability to bind to the ER in the absence of any of their metabolites. Comparison of IC_{50} and K_i values for the chemicals tested *in vitro* with that of endogenous and synthetic estrogens provide an indication of the potential of a given chemical to disrupt ER function *in vivo*. However, this assay does not distinguish between ER agonist and antagonists. The cytosolic rat ER binding assay may also yield false negative results if metabolic activation is required prior to binding to the ER or if the test chemical is not completely solubilized in the assay buffer. In addition, the results may be artifactual if ER is altered by detergent/denaturation effects of the test chemical, particularly if concentrations greater than 10 micromolar are used. At present, ER binding data are not entirely comparable from lab to lab because of methodological differences between labs in the conduct of this assay. However the rat cytosolic ER binding assay has been used for about 20 years, it is less complex than whole cell binding assays, and competent laboratories should be able to obtain similar results with minimal effort.

Cell-free and whole-cell binding assays using human ER (hER) are rapidly being developed and offer both advantages and disadvantages over the above assay, one advantage being the use of the human rather than the rat ER. However, being relative new, they have not been standardized in their examination of xenoestrogens. Assays for ER beta binding and/or transcriptional activation should be considered as they become more widely available, and included in screening if warranted (i.e., if it is determined that some xenobiotics bind only ER beta and would be missed in current assays with ER alpha).

a. ER Binding

The cell-free estrogen receptor alpha binding assay, a long-standing and relatively simple *in vitro* assay that detects specific mechanisms of endocrine activity, is recommended. This is important because several xenobiotics display affinity for the estrogen and/or androgen receptors. Binding assays identify, but do not discriminate between, agonists and antagonists. The apical nature of these assays is an advantage rather than a limitation because either activity can produce adverse reproductive effects. These assays typically lack metabolic activity, which is an advantage if one wishes to identify the specific compound with endocrine activity. However, the lack of metabolic activation is also a limitation because some xenobiotics require metabolic activation.

b. ER Transcriptional Activation

Binding of estrogen to ER α in target cells results in the initiation of specific transcription activation events. Various estrogen-regulated genes have been identified in MCF-7 cells

(pS2, Cath D, PgR, TPA), and their corresponding gene products can be measured as an endpoint for estrogen action (VanderKuur et al., 1993a; Pilat et al., 1993; Davis et al., 1995). However, such endogenous genes are additionally regulated by other cellular mechanisms (Nunez et al., 1989; Cavailles et al., 1989; Zacharewski et al., 1994), and the quantification of gene products (mRNA) may be relatively laborious and difficult. Therefore, the introduction of artificial, ER-regulated reporter gene constructs into MCF-7 cells has become a routine method of measuring ER transcriptional activation (VanderKuur et al., 1993b; Meyer et al., 1994). These reporter assays utilize the human ER of MCF-7 cells for transcriptional regulation of a reporter gene that codes for an exogenous enzyme that can be easily measured in a cell lysate. Of the typical reporter gene products of chloramphenicol acetyl transferase (CAT) and luciferase (Luc), the more sensitive assays utilize luciferase. Reporter genes can be introduced into cells for the duration of the experiment only (transient transfection) or permanently, creating a genetically altered subline (stable transfection).

Transcriptional activation assays are a direct manifestation of receptor-mediated responses on gene expression (i.e., the presence of a functional estrogen receptor and a reporter gene are sufficient to express estrogen-mediated induction). The **MVLN** assay (stably transfected MCF-7 cell line with an artificial gene including ER α, a controller segment of vitellogenin, and promotor regulating expression of luciferase), which detects transcriptional activation after receptor binding using a luciferase reporter gene, is recommended. This rapid and sensitive assay (IC50=20 pM range) confirms that ER binding and appropriate controls can distinguish agonists from antagonists. These assays should be conducted in a manner that allows them to detect receptor agonists as well. Although these assays often provide information similar to the above binding assays, this is not always the case, and there are well-founded biological reasons for a chemical to be positive in either the binding or the transcriptional activation assay but not both. However, due to a higher degree of difficulty, concern exists that proper execution of whole-cell assays requires a level of skill and training that may not currently exist in the toxicology community. If so, these assays might be much more difficult to implement than the binding assays, some of which have been used for decades and are less complex. In addition to the MVLN, other stably transfected cell lines have been or are being used to detect for ER and AR action.

In spite of the difficulty of establishing stably transfected cell lines, various MCF-7 cell derivatives have been created. The MVLN cell line is an MCF-7 cell derivative containing an artificial gene consisting of the ER-controlled segment of the vitellogenin promoter, regulating the expression of luciferase (Pons et al., 1990; Gagne et al., 1994). These cells also contain a neomycin resistance gene that was used in the stable transfectant selection process. Therefore, since all MVLN cells contain the reporter gene, estrogen-regulated transcription can be measured with a high sensitivity (E2 IC50=). However, the metabolic capability of the MVLN assay has not been studied in detail; it is assumed to be similar to that of MCF-7 cells from which they are derived. In principle, there are several advantages of this assay over other *in vitro* assays that assess estrogen action. The MVLN cell assay is easy to use because it is permanently transfected and it is a short-term assay. In addition, the MVLN cell assay has been standardized to the degree that is has been employed in high throughput transcription assays involving robotic manipulation of large numbers of sample wells

containing relatively few cells (e.g., 96 well plates). A procedure that has been used to characterize estrogen agonists as well as antagonists can be characterized with the MVLN assay (Gagne et al., 1994). In addition, a systematic comparison of more than 25 chemicals, including phthalates, alkylphenols, chlorinated pesticides, and steroids in the MVLN and the MCF-7 proliferation assay found that these assays were of equivalent sensitivity and responsiveness. Assays like the MVLN are deemed desirable because they are stably transfected and hence relatively easy to use and standardize, have high throughput potential, and are typically run to detect both agonists and antagonists.

The MVLN assay has been reported to have a disadvantage though, namely, that when the cells are briefly exposed to hydroxytamoxifen, their reporter gene cannot respond to estrogens. The mechanism underlying this effect is presently unknown. In principle, avoiding exposure to hydroxytamoxifen should prevent this from happening; however, this raises the issue of instability due to inadvertent exposure to chemicals during maintenance or propagation of the cells (this requires a serum-supplemented medium). The MVLN cells, like all other cell culture models, requires monitoring in order to ascertain that the initial response is preserved through extensive propagation (Badia et al., 1994).

2. Androgen Receptor Assays

a. AR Binding

The **cell-free AR binding** assay, used to determine the ability of environmental chemicals to compete with endogenous ligand for binding to AR, is recommended. This is an easy, time-honored task, with decades of use, and relatively simple to standardize and execute. Equilibrium binding assays require overnight incubation at 4^{o}C with AR isolated from castrated rat reproductive tissues (e.g., epididymis, ventral prostate, seminal vesicle) with increasing concentrations of radio-labeled ligand at different fixed concentrations of inhibitor or a fixed concentration of labeled androgen with increasing concentrations of unlabeled competitors. Following the incubation, hydroxyapatite or dextran-coated charcoal is used to separate protein-bound ligand from free ligand and specific binding is plotted in double reciprocal and as Scatchard plots as a function of competing inhibitor concentrations. Data analysis yields apparent equilibrium binding affinity constants for the inhibitor (Ki), which reflects the affinity of the chemical for the AR. Ki values can be used to rank chemicals for their ability to bind AR and potentially induce adverse effects. IC_{50} values can be used to calculate Ki values and the relative binding affinity (RBA) of the toxicant for AR, as compared to DHT or T, but this method is less accurate than experimental determination of the Ki. Within the last few years, a surprising number of chemicals in the environment from anthropogenic origin have been shown to act as AR ligands, including pesticides (e.g., vinclozolin, procymidone), pesticide metabolites (p,p' DDE and other DDT metabolites, methoxychlor metabolites), hydroxylated PCBs, and steroidal and non-steroidal natural and synthetic estrogens (Waller et al., 1996).

Advantages of the cell-free binding assay include ease of use, low cost, the potential to

standardize receptor preparations for distribution to many labs, and metabolism (but not spontaneous degradation) of chemicals in the assay is minimized. The absence of metabolism is an important consideration as parent chemicals and/or metabolites can be individually examined to determine which structure is responsible for AR binding, information that is critical if the data are to be used in a QSAR model. Disadvantages include the need for radio-labeled ligands and that data are restricted only to ligand binding affinity with no information on agonist or antagonist activity, AR stabilization, or degradation or rates of association and dissociation from the AR.

b. AR Transcriptional Activation

For AR-mediated activity, stably transfected cell lines are under development, but not yet widely available. The **AR transcriptional activation** (Cis-Trans) assay, using monkey kidney CV-1 cells, is recommended. A MCF-7 cell stably transfected with wild type androgen receptor has recently become available; however, only a few androgen agonists and antagonists have been tested using this cell proliferation assay (Szelei et al., 1997). Hence, like the CV-1, cell lines transiently cotransfected with hAR and a promoter construct with a Luc reported are recommended at this time. It is noteworthy that as compared to MCF-7 cells, the CV-1 has some metabolic capability. Here again, the YAS is not acceptable as it is unable to detect the AR-mediated activity of chlorinated pesticides.

Cells transiently transfected with hAR and reporter construct to detect transcriptional activation after receptor binding distinguish agonist/antagonist. Such assays have been used extensively and can be employed in a HTS mode for rapid screening. Transcriptional activation assays are used to determine whether chemicals which bind AR act as AR agonists or antagonists (Zhou et al., 1994; Simental et al., 1991). CV-1 cells are transiently transfected with the hAR expression vector together with a reporter construct (e.g., chloramphenicol acetyl transferase (CAT), beta-galactosidase, or firefly luciferase) containing an AR-dependent promoter such as the mouse mammary tumor virus promoter. Transfected cells are cultured in the presence (for antagonist activity) or absence (for agonist activity) of a single concentration of androgen (0.1 nM DHT) together with increasing concentrations of inhibitor. Following a 48h culture period, cells are harvested and luciferase activity is measured in the resultant solubilized cell extract as an estimate of AR-induced transcriptional activity.

Advantages of these types of assays are that they use human AR, they display some metabolic activity, and they establish whether a chemical that binds hAR acts as an agonist or antagonist. This information is critical in understanding the mechanism responsible for the induction of adverse endocrine-mediated developmental effects. Disadvantages of these assays are that they require the AR expression vector, reporter vectors, and transient cotransfections, which can be difficult. The assay requires close adherence to the standard operating procedure for reproducibility, and a 48h incubation during which time metabolism of the treatment chemicals may confound the data. In this regard, media from this assay, and other *in vitro* assays, should be analyzed before and after the incubation period to account for potential degradation and metabolism of the exogenous test chemicals and hormones.

3. Steroidogenesis

Antiandrogens and antiestrogens act via a number of direct mechanisms in addition to those that directly involve the steroid hormone receptors. One prominent mechanism of antihormonal activity is inhibition of hormone synthesis by inhibiting the activity of P450 enzymes in the steroid (and fungal) pathway. Such activity could be detected *in vitro* with a fairly simple *in vitro* procedure with minced testicular tissue obtained from adult male rats, because for many of the pesticides known to alter this pathway the parent material is active. Although aromatase, another P450 enzyme is present only at very low levels in the testis and male reproductive tract, it was proposed that inhibition of aromatase need not be included *in vitro* because it will be assessed in the *in vivo* pubertal female assay that follows. However, aromatase activity cannot be assessed in the proposed testis culture assay or in any of the *in vivo* assays using male rats.

The **testis culture *in vitro* assay using minced (50 mg) pieces of single testis,** which can be used to evaluate hormone synthesis with and without stimulation with cAMP, hCG, or substrates, is recommended. This assay assesses non-receptor mediated effects on P450 steroidogenic enzymes. Incomplete metabolism *in vitro* is of concern, exceptbut for those classes of chemicals where the parent material is active (e.g., certain classes of fungicides-drugs and agricultural products). This assay has been used for fetal, neonatal, and adult testis, and it is not limited to mammalian species, having been used to assess steroidogenesis in fish, reptiles, avian, and amphibian systems as well.

It is also possible to use cultures of Leydig cells isolated from testicular tissue to perform steroidogenesis assays. Leydig cells are the cells, within the testis, responsible for steroid synthesis. The advantage of using these isolates is that they are enriched for the cells that synthesize testosterone. The disadvantage is that there are extra steps in the preparation of the cells. Both approaches are expected to be comparable in their ability to detect steroidogenesis inhibitors. The utility of the minced testis culture is primarily based on data generated using Leydig cell cultures (Klinefelter and Kelce, 1996).

Substances that interfere with steroidogenesis act primarily by inhibiting cytochrome P450 enzymes in the steroid pathway. For example, the mechanism of action of two major classes of herbicidesfungicides, the imidazoles and the triazoles, involves inhibition of P450 enzymes in the sterol synthesis pathway for lanosterol, a vital component of fungal membranes (Talon et al. 1988). Cytochrome P450 inhibitors tend to be non-specific, and these fungicides can also inhibit other P450 enzymes such as those required for mammalian steroid hormone synthesis (Murray and Reidy, 1990). Inhibition of mammalian steroid synthesis can potentially result in a broad spectrum of adverse effects *in vivo*, including abnormal serum hormone levels, pregnancy loss, delayed parturition, demasculinization of male pups, lack of normal male and female mating behavior, altered estrous cyclicity, and altered reproductive organ weights.

D. Scientific Basis for *In Vivo* Screening for Estrogen, Androgen, and Thyroid Activities

The EDSTAC believes inclusion of *in vivo* methods in T1S can help reduce false negatives in the absence of knowledge of absorption, distribution, metabolism, and excretion. *In vivo* assays are often apical (that is, while they incorporate endocrine-specific endpoints, disruption of a number of hormone regulation/delivery mechanisms can be evaluated in the same assay). Therefore, they are less specific, but more comprehensive, than *in vitro* assays. *In vivo* assays can be made more specific if accompanied by target organ/cell dosimetry of biologically active metabolites. *In vitro* data are enhanced if the actual concentration of the chemicals in the media is determined, to account for metabolism, stability, and solubility, and to determine whether these concentrations compare to those that can be achieved *in vivo*. Cellular assays should determine viability, and the specificity and limitations of each assay should be defined. It is clear a combination of *in vivo* and *in vitro* assays is necessary in order to detect EAT alterations that act via the ER, AR, TR, inhibition of steroid hormone synthesis, and/or alterations of the hypothalamic-pituitary-gonadal (HPG) and HPT (thyroid) axis.

More than 50 assays, and related endpoints, were considered by the STWG, including *in vitro*, *in vivo*, and *ex vivo* (*in vivo* dosing followed by *in vitro* assessment of function) techniques. *In vivo* endpoints considered include reproductive organ weights and histology, serum hormone levels, *in vivo* gene activation, protein synthesis, behavior, growth, development, pregnancy maintenance, and anatomy/morphology. For each endpoint, the sensitivity (defined here as the response of the assay to low concentrations or dosage levels), specificity (pathognomonic for a mechanism of action, since the lack of specificity leads to false positives), relative simplicity, difficulties encountered running the assay, confounding factors, and limitations, test duration, and costs were discussed. In addition, items such as degree of acceptance of the method, how many chemicals had been screened, and the relative "newness" of the assay (state-of-the-art) were considered.

Advantages of *in vivo* assays include:

account for absorption, distribution, metabolism, and excretion;
well-defined, acceptable methods used for decades;
general acceptability in toxicity testing;
some endpoints are toxicologically relevant and have been used in risk assessment;
evaluate a broader range of mechanisms;
provide a comprehensive evaluation of the whole endocrine system as a unit; and,
give comparative perspective to other endpoints of toxicity.

It is important to reiterate that the screening battery is being designed to minimize false negatives, based on an assessment of the ability of the battery to detect known EDCs that act via EAT. In this regard, the value of each individual assay cannot be considered in isolation from the other assays in the battery, as they have been combined in a manner such that limitations of one assay are complemented by strengths of another.

The EDSTAC believes the proposed screening battery, once validated, will detect all of the EDCs mediated by EAT including xeno(anti)estrogens (that act via the ER or inhibition of aromatase by oral or parenteral administration), xeno(anti)androgens (via AR or hormone synthesis), altered HPG axis, and antithyroid action (via synthesis, metabolism and transport, and the TR). However, results of even the most specific *in vivo* assays can be affected by endocrine mechanisms other than those directly related to ER, AR, and TR action. For example, uterine weight in the ovariectomized female rat is affected in an estrogen-like manner by high doses of aromatizable and nonaromatizable androgens and growth factors like EGF. The age at puberty (vaginal opening in the female or preputial separation in the male rat) can be affected by chemicals that act on the hypothalamus, pituitary, or thyroid or alter growth hormone secretion. If gonadally intact females are used, uterine weight can also be affected by toxicants that stimulate hypothalamic-pituitary or gonadal endocrine secretions. Clearly, castration of the treated male or female markedly affects the specificity of the test. The lack of specificity of *in vivo* assays is a limitation if the goal is to only identify ER, AR, and TR alterations. In contrast, this lack of specificity could be considered an advantage if a broader, more apical screening strategy is desired.

1. Unique Thyroid Action Properties to be Considered in Design and Interpretation of T1S

Thyroid dysfunction leads to abnormal development, altered growth patterns, and a variety of physiological perturbations in mammals (Dussault and Ruel, 1987; Myant, 1971; Porterfield and Hendrich, 1993; Porterfield and Stein, 1994; Timiras and Nzekwe, 1989), as well as in birds (Tsai and Tsai, 1997), reptiles (Schrebier and Richardson, 1997), amphibians (Brown et al., 1995; Tata, 1994), and fish (Leatherland, 1994). Considering the consequences to wildlife populations and human health of the presence in the environment of synthetic compounds with thyroid disrupting activities, the EDSTAC has recommended a series of assays that will detect whether substances may interact with the thyroid.identify potential thyroid disrupting effects of such compounds based on existing information.

The chemistry of thyroid hormone, the endocrine mechanisms governing its regulation, and the mechanisms by which thyroid hormone exerts its effects are surprisingly similar among vertebrates (Gorbman et al., 1983). EDSTAC deliberations have, therefore, been guided by research focused on a variety of vertebrates to develop this series of screens. Despite the volume of literature reviewed, the current pace of research into thyroid hormone action makes it predictable that the present screens will become obsolete, both because more effective assays will likely be developed and because new information about thyroid hormone action may reveal mechanisms of thyroid activity disruption not identified by the proposed T1S battery. The following background information, about functioning within the thyroid axis and methods used to evaluate anti-thyroid activities, is intended to provide a rationale for the proposed Tier 1 antithyroid assays.

Endocrinology of the Vertebrate Thyroid:

Cells of the thyroid gland are arranged in follicles; the epithelial cells surround a fluid-filled core containing proteinaceous material - the colloid (Fawcett, 1986). Individual follicular cells respond to a pituitary hormone, thyrotropin (TSH), by increasing the synthesis and release of thyroid hormones (Wondisford et al., 1996). TSH release from the pituitary is stimulated, in turn, by a neuroendocrine peptide, thyrotropin-releasing hormone (TRH) (Greer et al., 1993; Morley, 1981; Taylor et al., 1990), and inhibited by the negative feedback effects of thyroid hormone itself (Franklin et al., 1987; Mirell et al., 1987; Shupnik and Ridgway, 1987). In a redundant negative-feedback loop, thyroid hormone also exerts an inhibitory effect on brain cells that manufacture TRH (Koller et al., 1987; Zoeller et al., 1993). The functional relationships among levels of this endocrine axis are so tightly linked that perturbations within one level produce compensatory changes in the other levels.

Thyroid Hormone Actions:

The majority of biological actions of thyroid hormones, including the regulation of brain development, are believed to be mediated by nuclear receptors for triiodothyronine (T3) (Lazar, 1993). Although the responsiveness to thyroid hormone requires the presence of nuclear TRs, the specific effects of thyroid hormone vary from tissue to tissue (Schwartz, 1983). Pleiotropic effects of thyroid hormone may be in part attributable to different levels and combinations of TR isoform expression (Lazar, 1993; Lazar, 1994). However, an important mechanism by which thyroid hormone effects can be regulated within cells, tissue, and across developmental stages is the interaction between receptors for thyroid hormone and those for retinoids (Forman and Samuels, 1990; Kliewer et al., 1992; Mano et al., 1994; Yu et al., 1991; Zhang et al., 1992). The implication of these observations is that thyroid hormone action can be modified, even disrupted, by interfering with retinoid metabolism.

Despite the recognition that thyroid hormone exerts its effects through nuclear receptors, there are very clearly defined endpoints of thyroid hormone action during development. There are a few genes expressed in mammals whose expression has been rigorously defined as directly regulated by thyroid hormone in the mammal. These include myelin basic protein (MBP) (Mitsubashi et al., 1988), neuroganin/RC3 (Iniguez et al., 1993), TRH (Hollenberg et al., 1995), malic enzyme (Song et al., 1986), thyrotropin (Carr et al., 1993), and some neuron-specific genes (Thompson, 1996). In amphibians, a number of genes have been identified in frogs (*Xenopus laevis*), which are shown to mediate effects of thyroid hormone on metamorphosis (Brown et al., 1996; Brown et al., 1995; Denver et al., 1997; Furlow et al., 1997; Kanamori and Brown, 1996), as well as more apical endpoints such as tailresorption.

Mechanisms of Antithyroid Activity:

A variety of environmental compounds are known to affect thyroid function or thyroid hormone action (Gaitan and Cooksey, 1989; Green, 1996). Processes known to be affected and a brief description of the effects are included below:

Active transport of iodide into the thyroid gland. Inhibitors include complex anions (e.g.,

ClO4, TcO4, thiocyanate). Reduces thyroid iodide uptake and can reduce thyroid hormone synthesis and circulating levels. Elevated TSH can overcome modest inhibition in the absence of other thyroid pathologies.

Iodination of thyroglobulin (by thyroid peroxidase). Inhibitors include thionamides (e.g., propylthiouracil, methiazole, carbimazole), thiocyanate, aniline derivatives such as sulfonamides, substituted phenols (resorcinol), flavonoids, and iodide. Reduces thyroid hormone synthesis and circulating levels, but can be overcome by elevated TSH.

Coupling reaction. Iodinated tyrosine residues of thyroglobulin must be coupled by an ether linkage to form iodothyronines, which are released from TG. Inhibition of this coupling reaction reduces thyroid hormone synthesis. This reaction may be controlled by TPO itself. Inhibitors include thionamides and other inhibitors of iodination, minocycline, lithium salts.

Hormone release. This is a cAMP-dependent process stimulated by TSH. Inhibitors include iodide and lithium salts.

Iodotyrosine deiodination. This process is important for recovery of iodide within the thyroid gland. Inhibition causes the reduction in thyroid iodide content and thus, inhibition of thyroid hormone synthesis. Inhibitors include nitrotyrosines.

Iodothyronine deiodination. This reaction is important for conversion of thyroxine to the hormonally active tri-iodothyronine, and for the conversion of T3 to the hormonally inactive T2. Inhibitors include thiouracil derivatives, oral cholecystographic agents, amiodarone.

Hormone excretion or inactivation. Inducers of hepatic drug-metabolizing enzymes. Inhibitors include phenobarbital, phenytoin, carbamazepine, rifampicin, organochlorines.

Hormone action. Thyroid hormone action is largely mediated by binding to specific nuclear receptors. There is limited evidence that compounds such as phenytoin (dilantin) and amiodarone can displace T3 from nuclear binding sites *in vitro*. However, there is little *in vivo* evidence that this interaction may compromise T3 action. In addition, there are predictions that specific PCBs may interfere with T3 binding to the nuclear receptor because of similarities in structure. However, these predictions have not been experimentally verified as yet.

Proposed T1S Assays for Anti-thyroid Activity:

To the EDSTAC's knowledge, all known antithyroid compounds so far reported affect circulating levels of thyroxine (T4). The physiological consequences of these affects are variable and may require considerable time to develop in a screening paradigm. In addition, they may represent endpoint measures that are not solely responsive to thyroid disruption. Therefore, the proposed T1S battery includes *in vivo* measures of circulating levels of thyroxine and TSH, and changes in the histopathology of the thyroid gland, as initial endpoints for antithyroid screening. These measures can be taken in animals that are treated in the commission of other *in vivo* screens (e.g., uterotrophic assay). Commercial kits are available to measure T4 and TSH by radio-immunoassay, and these assays are validated and standardized. It is important to recognize that a significant change in circulating thyroxine or TSH should be considered a positive finding. This decision is based upon the fact that some compounds, such as PCBs, have been reported to reduce circulating levels of T4, but leave

TSH unaffected (Goldey et al., 1995). In addition, weak antithyroid agents, especially those affecting some aspects of iodide metabolism in the thyroid (Gaitan and Cooksey, 1989; Gaitan et al., 1989; Green, 1996), may be compensated for by elevated TSH. Thus, T4 may appear normal, but TSH would be elevated. Finally, the absence of an effect on circulating levels of T4 or TSH does not preclude the possibility that an agent is antithyroid. It is well known that goitrogens can affect thyroid function over long periods and not be manifested by significant changes in circulating levels of T4 or TSH measured by radio-immunoassay (Gaitan et al., 1989). These compounds would produce a measurable effect on the thyroid gland. For these reasons, the EDSTAC recommends analysis of thyroid histopathology.

During their deliberations, the STWG extensively discussed the timing of exposure to a CSM. EDSTAC is recommending evaluation of antithyroid effects in animals prepared for testing other actions (either 14-day or 20-day exposure). Although cases in which exposure to xenobiotics of greater than 14 days are required to significantly affect circulating levels of T4, TSH, or thyroid histopathology are unknown, EDSTAC believes longer periods may be required. Duration of CSM exposure must be quickly evaluated in the validation phase.

These measures in mammals represent evaluation of thyroid function; there are no clear markers of thyroid hormone action that could be used within the context of a T1S assay. In contrast, tail resorption in amphibian metamorphosis represents an assay which utilizes specific thyroid hormone-dependent effects as an endpoint for a T1S assay.

2. *In Vivo* Assays Using Other Vertebrates

The T1S battery includes an amphibian and a fish assay, which fill important needs in the battery and complements the information from assays using mammals. These assays help the battery meet design criteria 2 and 4, which express the need for a sufficient range of taxonomic subjects and range of metabolic functions be evaluated in the battery. While the basic biochemical processes of receptor binding and cellular activation by hormones are known to be similar among many organisms, detailed comparative data do not exist to assess the extent of the homology across vertebrate classes. In particular, fish ER differs from mammalian ER more than the ER of other classes, and fish have some unique androgens. Hence, including fish as subjects makes sense as it is the class most likely to show differences from mammals in EAT activity. In addition, there are known differences in the ability of organisms to metabolize xenobiotics, due partially to the route of exposure. Fish and amphibians receive more dermal exposure to CSMs than other vertebrate classes, and thus CSMs avoid immediate metabolism in the liver. Theses assays also meet the need for the battery to have a clear cut response to measure the effects of thyroid hormone, which frog metamorphosis does, and hence complements the data on serum concentration of thyroid hormones and thyroid gland histology derived from the pubertal female rodent assay.

Unlike the mammalian assays, the assays recommended for fish and frogs have not been used in regulatory testing, and hence they need more work before being implemented for that purpose. In fact, however, both procedures using these species and similar endpoints have been used to investigate endocrine activity of CSMs in the research literature. The work needed for standardization of protocols and validation of the assay with known endocrine disruptors should proceed as soon as possible as these assays play a crucial role in the T1S battery. In addition, EDSTAC encourages development of other assays in the event that either of these two fail to be adequately standardized and validated, so that a complete screening battery can be implemented.

E. *In Vivo* Assay Overviews

Several measures of estrogenicity (reviewed by Gray et al., 1997; Reel et al., 1996; Parker, 1966, Chapter 30) have been used for over 70 years, including uterine size, vaginal cornification, female sexual receptivity, and age at puberty/vaginal opening (see Marshalls Physiology of Reproduction, 1966, for a thorough review). For example, Cook et al. (1938) found that DES produced full estrus in ovariectomized rats, so far as vaginal, uterine, and mating reactions were concerned. These remain some of the most useful short-term *in vivo* methods for screening for estrogenicity. Studies of xenoestrogens typically indicate that the sensitivity of these endpoints is as follows: uterine weight measured 5 hours after the last treatment, with fluid, is generally more sensitive than the age at vaginal opening or vaginal cornification; however, this is not always the case. Uterine histology and biochemical measures appear to be at least as sensitive to estrogens as uterine weight, but these endpoints are slightly more difficult to evaluate as they require specialized skills and equipment and are more expensive.

The sensitivity of the age at vaginal opening to methoxychlor appears to be about twofold greater than the onset of persistent vaginal cornification (PVC) and at least equivalent to the sensitivity of the uterotrophic assay (Gray et al., 1989; Gray and Ostby, in press). However, in another study hydroxylated PCBs induced vaginal cornification at dosage levels that failed to induce an increase in uterine weight (Gillesby and Zacharewski, 1996). PVC was not detected in a longterm study of the estrogenicity of octylphenol that doubled uterine weight in long-term ovariectomized rats after 10 weeks of oral administration and after three days of administration in juvenile rats (Gray and Ostby, in press). Hence, some of the original measures of estrogenicity, in use now for nearly three-quarters of a century, are still regarded as the most useful indicators of estrogenic activity *in vivo*.

Recent studies to evaluate methoxychlor, 4-tert-octylphenol, nonylphenol, bisphenol A, DES (Reel et al., 1985), estrogens, and antiestrogens (Conner et al., 1996) have demonstrated the utility of these biological assays, since: 1) pro-estrogens/metabolites may be detected following *in vivo* exposure; 2) agonistic/antagonistic properties may be addressed; 3) bioaccumulation and/or the development of tolerance to exposure may be evaluated; 4) multiple routes and lengths of exposure may be easily compared; and 5) acute exposure

regimes may be used. However, care should be taken when interpreting results from these biological assays since: 1) some environmental chemicals do not test positive for all measures; 2) exposure route and time of assessment following exposure affect the results; and 3) the observed biological response may result from other mechanisms of action. For example, in the ovariectomized female, increased uterine weight can be induced by aromatizable and nonaromatizable androgens (Salamon, 1938) and EGF (Nelson et al., 1991).

In the intact (ovaries present versus castrate or ovariectomized) juvenile female rat, the age at which treatment is initiated (typically 19-21 days of age) and the duration of treatment are critical variables that affect uterine weight. Exposure duration longer than 3 or 4 days or the use of juvenile females 24-25 days of age at the start of the study are not recommended because of the potential confounding of the treatment effect with the onset of natural estrous cyclicity and its concurrent fluctuations in uterine weight and histology. As long as the uterine weight bioassay has been used, it still has not been completely standardized, a fact that leads to some variation in results from lab to lab. For example, there are differences with respect to how well the mesenteric fat along the uterine circulation is trimmed, and some labs weight the uterus with its contents, while others remove the fluid before weighing. Uterine weight, serum hormone concentrations, and other evaluations in intact female rats are difficult to interpret unless great care is taken to assure that females are necropsied at the same stage of the estrous cycle. With regard to the measurement of serum hormones in the cycling female rat, the time of day is also critical, in addition to the day of the cycle. Effects on estrous cyclicity are not limited to ER-mediated alterations; several other reproductive (hypothalamic-pituitary) and nonreproductive (hypothyroidism) endocrine-related alterations can alter estrous cyclicity in the female rodent. The detection of vaginal cornification in juvenile, and ovariectomized adult rodents is one of the original assays used to detect estrogenicity and, as indicated above, this assay appears to be relatively sensitive to weak estrogens. However, higher levels of xenoestrogens are required to disrupt estrous cyclicity and induce constant estrus persistent vaginal cornification in intact female rats (Gray et al., 1989).

1. Rodent 3-Day Uterotrophic Assay (Subcutaneous)

Assay for estrogenicity

An increase in uterine weight is generally considered to be one of the Gold Standards of Estrogenicity when measured in the ovariectomized (ovx) or immature female rat or mouse after 1-3 days of treatment. The recommended 3-day uterotrophic assay (sc injection) uses the ovariectomized adult female rat (the duration can be extended if so desired) with n =10/group. Sc- treatment is recommended at this time because most of the historical data are collected in this manner and there is relatively little data concerning the effects of other routes of administration at this time. At necropsy one should carefully trim the uterus of fat and weigh with and without fluid and save uterus and vaginal tissues for histopathology. Most xenoestrogens have been examined in this assay. It also should be executed in a manner to detect antiestrogens. In this regard, a control and xenobiotic-treated group are

coadministered estradiol sc and necropsied.

2. Rodent 20-Day Pubertal Female Assay with Thyroid

Assay for thyroid, HPG axis, aromatase, and estrogens that are only effective orally or after longer dosing than the uterotrophic assay.

The determination of the ages at "puberty" in the female and male rat are endpoints that already have gained acceptance in the toxicology community. Vaginal opening in the female and preputial separation in the male are required endpoints measured in the new EPA multigenerational test guidelines. In this regard, this assay would be easy to implement because these endpoints have been standardized and validated and VO and PPS data are currently being collected under GLP conditions in most toxicology laboratories. In addition, VO and PPS data are reported in many recently published developmental reproduction studies (i.e., see studies from R.E. Petersons, R. Chapins and L.E. Grays laboratories on dioxins, PCBs, antiandrogens, and xenoestrogens).

In the pubertal female assay, oral dosing is initiated in weanling rats at 21 days of age (10 per group, selected for uniform body weights at weaning to reduce variance). Dose daily, 7 days a week, and examine daily for vaginal opening (could also check for age at first estrus and onset of estrous cyclicity). Dose until vaginal opening is attained in all females (typically two weeks after weaning, unless delayed). Determine age at vaginal opening (VO) in female rat. Rats are dosed by gavage with xenobiotic and examined daily for VO. Advantage over uterotrophic assay is that one test detects both agonists and antagonists, it detects xenoestrogens like methoxychlor that are almost inactive via sc injection, it detects aromatase inhibitors, altered HPG function, unusual chemicals like betasitosterol. In addition, at necropsy weigh ovary (increased in size with aromatase inhibitors, but reduced with betasitosterol), save thyroid for histopathology, take serum for T4, and measure TSH.

Exposure of weanling female rats to environmental estrogens can result in alterations of pubertal development (Ramirez and Sawyer, 1964). Exposure to a weakly estrogenic pesticide after weaning and through puberty induces pseudoprecocious puberty (accelerated vaginal opening without an effect on the onset of estrous cyclicity) after only a few days of exposure (Gray et al., 1989). Pubertal alterations also result in girls exposed to estrogen-containing creams or drugs, which induce pseudoprecocious puberty and alterations of bone development (Hannon et al., 1987).

Several examples of estrogenic chemicals affecting vaginal opening in rodents are known and include methoxychlor (Gray et al., 1989), nonylpheno, and octylphenol (Gray et al., in press). This endpoint appears to be almost as sensitive as the uterine weight bioassay, but the evaluation is easier to conduct and does not require that the animals be euthanized, so they can be used for additional evaluations. For example, treatment with methoxychlor at weaning (6 mg/kg/d or higher) caused pseudoprecocious puberty in female rats. Vaginal opening occurs from two to seven days earlier in treated animals than controls, in a dose-related

fashion, but methoxychlor did not alter estrous cyclicity at the low dosage levels, indicating a direct estrogenic effect of methoxychlor on vaginal epithelial cell function without an affect on hypothalamic-pituitary maturation. Similar effects have been achieved with chlordecone, another weakly estrogenic pesticide, and octylphenol. Chlordecone also induces neurotoxic effects (hyperactivity to handling and tremors). In addition to estrogens, the age at vaginal opening and uterine growth can be affected by alteration of several other endocrine mechanisms, including alterations of the hypothalamic-pituitary-gonadal axis (Shaban and Terranova, 1986; Gonzalez et al., 1983). In rats, this event can also be induced by androgens (Salamon, 1938) and EGF (Nelson et al., 1991). In the last 20 years there have been over 200 publications which demonstrate the broad utility of this assay to identify altered estrogen synthesis, ER action, growth hormone, prolactin, FSH or LH secretion, or CNS lesions.

3. Rodent 5-7 Day Hershberger Assay

Assay for Antiandrogens

In the castrated male rat, the gonads have been removed and effects on androgen-dependent tissues are likely to be direct and not a result of pituitary or gonadal secretion. The assay (Hershberger, 1953) requires two stages as below:

castrated male rat + T + Xenobiotic
castrated male + X (to detect agonist)

In this *in vivo* test, sex accessory gland weights (ventral prostate and seminal vesicle separately) are measured in castrated, testosterone-treated adult male rats after 4-7 days of treatment by gavage with the test compound. The advantage of this assay is that it is fairly simple, short term, and relatively specific compared to other *in vivo* procedures. Although the androgens, testosterone and DHT, play a predominant role in the growth and maintenance of the size of these structures, several other hormones and growth factors can influence sex organ weights including the thyroid and growth hormones, prolactin and EGF (Luke and Coffey, 1994). Exposure to estrogenic pesticides can also reduce sex accessory gland size; however, it is unclear to what degree these reductions result from direct versus indirect action of the chemical. Other useful endpoints that help reveal the mechanism of action include serum hormone levels of T, DHT, LH, AR distribution, TRPM2/C3 gene activation, ODC, and 5α reductase activity in the prostate. The prostate and seminal vesicles should be weighed separately because these organs differ with respect to the androgen that controls their growth and differentiation. The prostate is dependent upon enzymatic activation of T to DHT, whereas the seminal vesicle is less dependent upon this conversion. Hence, effects on 5I reductase can be distinguished from AR-mediated mechanisms by determining whether the prostate is preferentially affected. Growth of the levator ani muscle is T dependent, having little capacity to convert T to the more potent androgen DHT. Weight of this muscle is useful in identifying anabolic androgens and antiandrogens, and for this reason has been used extensively in the pharmaceutical industry. In order to detect androgenic rather than antiandrogen action one would simply delete the hormone administration from the protocol.

Data from this assay (often with slight modifications), using drugs and xenoantiandrogens, are widely available in the literature. For a *non-in utero* assay, this assay robustly detects androgens and antiandrogens with a dynamic response that typically exceeds that of the intact adult male (Raynoud, 1984). Most of the studies are able to detect significant effects with only five animals per group. In fact, in one study which used 10-15 drugs, the Hershberger assay was more responsive than was the intact male for every chemical examined (Wakeling et al., 1981). The power of the castrate-male assay arises from the fact that castration creates a "fetal-like" endocrine system with regard to the regulation of androgen secretion, as the HPG axis can no longer compensate for the effect of the chemical on the AR. For example, p,p' DDE reduces sex accessory gland weights in this assay, but not when administered to intact male rats (Kelce et al., 1996; 1997).

4. Frog Metamorphosis Assay

This assay employs intact larval (tadpole) stages of the African clawed frog *(Xenopus laevis)* exposed over a 14-day time period, 50-64 days of age, to observe the rate of tail resorption (Fort and Stover, 1997). Tail resorption can be easily quantified with computer-aided video image processing (Fort and Stover, 1997). The molecular mechanisms involved in tail resorption are well characterized (Brown et al., 1995; Hayes, 1997a) and this assay is, therefore, considered to be a simple and specific assay for thyroid action. It will detect thyroid (increase in tail resorption rate) and antithyroid (decrease in tail resorption rate) effects. Because evidence also suggests that thyroid action on tail resorption is regulated by corticoids, estrogens, and prolactin (Hayes, 1997b), this assay will address distinctive modulating pathways and, in tandem with the 14-day mammalian pubertal assay, a comprehensive screen for thyroid hormone activity is achieved.

5. Fish Gonadal Recrudescence Assay

Intact mature fish maintained under simulated "winter" conditions (short day length, cool temperatures) exhibit regressed secondary sex characteristics and gonad maturation. In this assay, intact fish of both sexes (fathead minnow, *Pimephales promelas*, or other appropriate speciesrecommended) are simultaneously subjected to an increasing photoperiod/temperature regime and test substance to determine potential effects on maturation from the regressed position (recrudescence). The primary endpoints examined in the assay include morphological development of secondary sexual characteristics, ovary and testis development (weight increases), gonadosomatic index (ratio of gonadal weight to body weight), final gamete maturation (ovulation, spermiation), and induction of vitellogenin. This assay is sensitive to BPG axis effects in addition to androgen- and estrogen-related activity.

Fish differ in steroid profiles from mammals; for example, 11-ketotestosterone as opposed to testosterone is the most important androgen in fish. The estrogen receptor in fish appears to differ structurally and functionally from the mammalian estrogen receptor (Petit et al., 1995; Gustafsson, 1996). Also, steroid receptors in eggs and for vitellogenin have no known

analogous receptors in mammals, which would suggest sites of endocrine disruption unique to oviparous animals. Therefore, this assay is essential to address these known endocrine differences.

F. Alternative Assays for Possible Inclusion

1. Placental Aromatase Assay

One critical enzyme present at very low levels in the testis, and at higher levels in the ovary, uterus, and placenta, is aromatase, which converts testosterone to estradiol and is another P450 isozyme. Human placental aromatase is commercially available and could be used *in vitro* to assess the effects of toxicants on this enzyme fairly easily.

2. Modified Rodent 3-Day Uterotrophic Assay (Intraperitoneal)

This is an *in vivo* assay (O'Conner et al., 1996) for estrogenic activity in ovariectomized female rats. It can detect certain antiestrogens with mixed activity, i.e., some agonistic activity (e.g., tamoxifen). The rats are injected intraperitoneally with the test agent daily for three days. The ip injection method may enhance the sensitivity of the assay and is capable of detecting the estrogenic potential of methoxychlor, which has been cited as an example of a compound not detectable by the sc route. The females are necropsied either 6 hours or 24 hours after the final treatment, depending on the protocol employed by the laboratory. Vaginal cytology is evaluated by vaginal lavage to determine whether the epithelium has become cornified, indicative of estrous. Presence of fluid in the uterine lumen is noted and recorded, and the number of animals that have fluid in the uterus is reported. Fluid imbibition is indicative of estrogenic potential. The uterus is excised and weighed. It is then preserved in an appropriate fixative for subsequent histological evaluation, if needed.

Subsequent histological evaluation will be triggered by an equivocal uterine weight or uterine fluid response (i.e., an increase that is not statistically significant). This evaluation will consist of a characterization of the appearance of the uterine epithelium, a measurement of uterine epithelial cell height, and epithelial mitotic index or proliferating cell nuclear antigen (PCNA) immunohistochemistry. Uterine cell height and cell proliferation are sensitive indicators of estrogenic potential.

3. 14-Day Intact Adult Male Assay

This *in vivo* assay is intended to detect effects on male reproductive organs that are sensitive to antiandrogens and agents that inhibit testosterone synthesis or inhibit 5-alpha-reductase (Cook et al., 1997). The duration of the assay is anticipated to be sufficient to detect effects on thyroid gland activity. The rats are anatomically intact and mature; therefore, they have

an intact HPG axis, allowing an assessment of the higher order neuroendocrine control of male reproductive function and the thyroid.

Young adult male rats (70-90 days of age) are used in this assay. They are dosed daily with the test agent for 14 days. The recommended route of administration is ip, which may, in some cases, maximize the sensitivity of the assay. They are necropsied 24 hours after the final dose. Immediately after sacrifice one cauda epididymis is weighed and processed for evaluation of sperm motility and concentration. The following organs are weighed: testes, epididymides, seminal vesicles, and prostate. The following are fixed and evaluated histologically: one testis and epididymis, and the thyroid. The following hormones are measured in blood plasma: T4, TSH, LH, testosterone, DHT, and estradiol.

Empirical assessment of this assay has shown it to be sensitive to agents that are directly antiandrogenic, inhibit 5-alpha-reductase, inhibit testosterone synthesis, or affect thyroid function. The sensitivity of this assay, as defined as the ability to detect a hazard, may be comparable to other assays that have been proposed.

4. Rodent 20-Day Thyroid/Pubertal Male Assay

This assay detects androgens and antiandrogens *in vivo* in a single stage apical test. "Puberty" is measured in male rats by determining age at PPS (preputial separation). Animals are dosed by gavage beginning one week before puberty (which occurs at about 40 days of age) and PPS is measured. Androgens will accelerate and antiandrogens and estrogens will delay PPS. The assay takes about 3 weeks, is easy to determine PPS, and allows for comprehensive assessment of the entire endocrine system in one study (10 per group, selected for uniform body weights to reduce variance). Dose daily, seven days a week, and examine daily for PPS. Continue dosing until 53 days of age and necropsy males. Weigh body, heart (thyroid), adrenal, testis, seminal vesicle plus coagulating glands (with fluid), ventral prostate, and levator ani plus bulbocavernosus muscles (as a unit). Save the thyroid for histopathology and take serum for T4, T3, and TSH. Testosterone, LH, prolactin, and dihydrotestosterone analyses are optional. These endpoints take several weeks to evaluate and are affected not only by estrogens but by environmental antiandrogens, drugs that affect the hypothalamic-pituitary axis (Hostetter and Piacsek, 1977; Ramaley and Phares, 1983), and by prenatal exposure to TCDD (Gray et al., 1995a; Bjerke and Peterson, 1994) or dioxin-like PCBs (Gray et al., 1995b). In contrast to these other mechanisms, only peripubertal estrogen administration accelerates this process in the female and delays it in the male. Preputial separation in the male rodent is easy to measure and this is not a terminal measure (Korenbrot et al., 1977).

Age and weight at puberty, reproductive organ weights, and serum hormone levels can also be measured. Delays in male puberty result from exposure to both estrogenic and antiandrogenic chemicals including methoxychlor (Gray et al., 1989), vinclozolin (Anderson et al., 1995) and p,p' DDE (Kelce et al., 1995). Exposing weanling male rats to the antiandrogenic pesticides p,p' DDE or vinclozolin delays pubertal development in weanling

male rats as indicated by delayed preputial separation and increased body weight (because they are older and larger) at puberty. In contrast to the delays associated with exposure to estrogenic substances, antiandrogens do not inhibit food consumption or retard growth (Anderson et al., 1995). Antiandrogens cause a delay in preputial separation and affect a number of endocrine and morphological parameters including reduced seminal vesicle, ventral prostate, and epididymal weights. It is apparent that PPS is more sensitive than are organ weights in this assays. In addition, responses of the HPG are variable. In studies of vinclozolin, increases in serum LH were a sensitive response to this antiandrogen, whereas serum LH is not increased in males exposed to p,p' DDE during puberty (Kelce et al., 1997). Furthermore, a systematic review of the literature indicates that the sex accessory glands of the immature intact male rat are consistently more affected than in the adult intact male rat.

In summary, preputial separation and sex accessory gland weights are sensitive endpoints. However, a delay in preputial separation is not pathognomonic for antiandrogens. Pubertal alterations result from chemicals that disrupt hypothalamic-pituitary function (Huhtaniemi et al., 1986), and, for this reason, additional *in vivo* and *in vitro* tests are needed to identify the mechanism of action responsible for the pubertal alterations. For example, alterations of prolactin, growth hormone, gonadotrophin (LH and FSH) secretion, or hypothalamic lesions alter the rate of pubertal maturation in weanling rats.

As indicated above, the determination of the age at "puberty" in the male rat are endpoints that already have gained acceptance in the toxicology community. Preputial separation in the male is a required endpoint in the new EPA multigenerational test guidelines. In this regard, this assay would be easy to implement because these endpoints have been standardized and validated and PPS data are currently being collected under GLP conditions in most toxicology laboratories. In addition, PPS data are reported in many recently published developmental reproduction studies (i.e., see studies from R.E. Petersons, J. Ashbys, R. Chapins and L.E. Grays laboratories on dioxins, PCBs, antiandrogens, and xenoestrogens).

Sex accessory gland weights in intact adult male rats also can be affected directly or indirectly by toxicant exposure. The HPG axis in an intact animal is able to compensate for the action of antiandrogens by increasing hormone production, which counteracts the effect of the antiandrogen on the tract (Raynoud et al., 1984; Edgren, 1994; Hershberger, 1953).

IV. General Principles in Evaluating Tier 1 and Tier 2 Results

A. Introduction

Substances screened for hormonal activity and/or tested to determine whether any endocrine-mediated adverse effects occur for their potential to be endocrine disruptors will be subjected to multiple assays, both *in vitro* and *in vivo*, in the Tier 1 and Tier 2 phases. Apart from substances yielding negative results in all assays, it is likely that most substances will produce a unique array of results requiring a judgment as to whether the weight of evidence indicates

the substance should or should not be judged a candidate for T2T (after completing T1S), classified first, as a potential endocrine disruptor (in T1S), and second, designated as an endocrine disruptor (after completing in T2T). A table consisting of 18 chemical types along with known or expected T1S results can be found in Appendix O, Examples of "Weight of Evidence" Determinations.

There are two senses in which a "weight of evidence" determination will need to be made. The first is with respect to the question of whether consistent results are being obtained across multiple assays. If the results are not consistent, it will be necessary to "weight" the conflicting results, allowing some to carry more weight than others. The second sense is with respect to the question of whether a particular body of evidence, even if it is fully consistent, is sufficient to justify a decision. In this sense, it is the "weight" of the entire body of evidence, relative to some minimal level established as being required for sound decisions, that is being judged.

Assessing the "weight-of-evidence," and using that assessment in forming judgments about a to judge the classification of a substance, can be done in a variety of ways. On one extreme are approaches based solely on expert judgment in which an individual reflects on the data and offers an informed, yet a personal, opinion as to the classification. On the other extreme are more formal and mathematical procedures such as Bayesian analysis in which data are viewed sequentially and used to formulate prior and posterior judgments. An intermediate approach is one in which a group debates the available data, presents alternative arguments for the classification, and collectively reaches a judgment.

All three of these possibilities are forms of "weight-of-evidence" assessments. The EDSTAC has agreed not to prescribe a particular "weight-of-evidence" approach, as these are controversial and a matter of science policy to be established by the Agency. Instead, we offer here general guidelines for reasoning from the data produced in the two tiers, which conform to the outline provided in the NAS/NRC report Science and Judgment in Risk Assessment (National Academy Press, 1994). These guidelines provide a framework within which one may take into account multiple features of screening and testing data as these are relevant in determining whether the substance should be a priority for T2T is a potential endocrine disruptor (after T1S) and/or is determined to have endocrine disrupting effects should be classified as an endocrine disruptor (after T2T).

"Weight-of-evidence" considerations will arise at two places within the EDSTP. It first will arise in considering whether the body of evidence collected solely within a given tier (e.g., Tier 1) warrants a particular conclusion (e.g., that the substance may have does or does not have the potential for endocrine activitydisruption). The second place where it will arise is in considering whether results from a previous tier (e.g., Tier 1) should affect the conclusions drawn from the a subsequent tier (e.g., Tier 2). By this, we are not referring to the fact that Tier 1 results may guide selection and/or design of Tier 2 Tests (with the results of the Tier 2 Tests then being interpreted without further reference to the Tier 1 results). We are, instead, referring to the possibility that the results of the T1S assays may be "weighted into" the determination of whether a substance has passed or failed the Tier 2 Tests.

A broad range of results may need to be weighted into a final judgment at either tier. Information routinely taken into consideration in determining the "weight-of-evidence" will include:

the balance of assays/tests that gave positive and negative results;
results of *in vitro* versus *in vivo* assays/tests;
the nature of the biological effects induced;
the range of effects observed;
the slope and shape of the dose-response curves;
the level, magnitude, or severity of the effects induced; and
the presence or absence of response in multiple taxa.

The "weight-of-evidence" approach makes explicit the assumption that results of some assays/tests, in some taxa, at some level of severity (etc.) are intrinsically "worth" more than others and should, therefore, carry more weight in the final decisions following T1S and T2T. For example, positive results showing reproducible, high levels of effects at low doses (near the doses produced by environmental or human exposures) are likely of greater weight than weak effects observed only at very high, perhaps toxic, levels of exposure.

The EDSTAC has taken the approach here of providing guidance on the use of "weight-of-evidence" by advising that any approach used must satisfy several broad criteria which we take to be essential. Some of these are general criteria. The weighting system should be transparent, allowing individuals to review the "weight-of-evidence" determination. It should be possible to understand the procedure before viewing the data, so individuals have a reasonable expectation of the final decision at the time when the data are presented. This does not mean the decision is fully determined by the data, removing the need for scientific judgment, but it does mean that any deviations from the expected decision should be supported by an explanation detailing the "weight-of-evidence" assigned.

B. False Negatives and False Positives Within the Context of T1S and T2T

The guiding principle for the treatment of false positives and negatives should be one of valuing sensitivity more than specificityerring on the side of false positives unless this compromises the ability to sort CSMs into a subset most likely to be of concern. False positives and negatives can arise in at least three different ways in the screening and testing batteries (see Figure 5.1):

The false result may be due to the stochastic nature of screens and tests. A false result leading to an incorrect claim that the screen/test is positive is a Type I (false positive) error. A false result leading to an incorrect claim that the screen/test is negative is a Type II (false negative) error. The frequency of these types of errors is expressed by the p value for an assay, so the selection of a required p value to classify a result as positive will determine the frequency of Type I and Type II errors. The guiding principle above suggests that required p values should be chosen so Type II errors are minimized, while

also ensuring that Type I errors do not become so frequent that CSMs can no longer be sorted meaningfully.

False positives and negatives may arise due to unknown or unexpected limitations of the test or assays, such as anomalous activity of chemicals or classes in a particular assay or interference from assay procedures.

The third source of error arises from a potential lack of predictivity of results in T1S for endocrine disruptive responses in T2T. This source of error is shown in Figure 5.1 by the bold arrow going from positive results in T1S to T2T. A negative result in T1S may simply mean the assay battery misses a mechanism of action that would have been active in a T2 Test. This will result in a false classification of the substance as not having endocrine activitydisrupting potential, an error that would have been caught at T2 had the CSM proceeded to that stage. For this reason, the T1S battery was designed to capture all known endocrine mechanisms for EAT and to minimize false negative results specifically as opposed to false positives. A positive result in T1S could be followed by negative results in T2T because the endpoints measured in T1S may not accurately predict adverse effects in long-term, whole animal tests. This will result in unnecessary testing of some chemicals in T2T, a possibility considered more acceptable than missing potential endocrine disruptors.

In treating the frequency of Type I and Type II errors, it is important to consider both the frequency of these errors in each particular assay/test and the number of assays/tests in a battery. As the number of assays/tests in a battery increases, the probability that AT LEAST ONE of the assays/tests will show a false positive increases. This is shown in Figure 5.2, which displays the relationship between the probability of any one assay/test showing a false positive (the X axis), the number of assays/tests in the battery (the Y axis), and the probability that the battery shows at least one false positive result. In this figure, it is assumed that the CSM tested actually has no endocrine activitydisrupting potential, but might yield a false positive result due primarily to stochastic variation. The goal of the Tier 1 or Tier 2 stages should be to minimize the probability of a battery producing a Type II error for a CSM, while not causing the probability of a Type I error from getting so large that the battery becomes ineffective at sorting CSMs. Figure 5.3 displays the analogue of Figure 5.2; i.e., the probability that a battery produces a false negative result if each assay/test has a given false negative frequency, there are N assays/tests in the battery, and the CSM truly is an endocrine disruptor.

We caution that the statistical properties of actual assays/tests in a battery will not be identical, so Figures 5.2 and 5.3 are simply illustrative and must be modified for any particular battery developed. What the figures indicate is simply that the decision as to how to weight a single positive result from a battery into the "weight-of-evidence" judgment should reflect a concern for both Type I and Type II errors. From these figures, it can be seen that a large battery (e.g., with 10 assays/tests), each with a false positive frequency of only 10%, can result in a very high probability of producing at least one assay/test showing a false positive when applied to a substance that in reality has no endocrine-disrupting properties. Such a

battery would be essentially useless in sorting CSMs and focusing society's resources. Our final advice here is that an effort should be made to characterize statistically the frequency of Type I and Type II errors associated with any selected battery, and to use this characterization in deciding the weight assigned to a single positive result from that battery.

C. Specific Principles for Evaluating T1S

There are several specific criteria to be met by the decision process assuming appropriate dose and route of exposure as discussed previously in the Chapter:

1. If equivalent information is available (e.g., from the sorting and prioritization phase), it may be that only those T1S assays which evaluate the endocrine activitydisruption of potential concern from a CSM would be performed, i.e., only a subset of assays would be run. Similarly, the results of the T1S assays may require that only a subset of the T2 Tests be conducted.

2. If all assays are performed, and all assays are negative, then the CSM does not have endocrine activity for estrogen, androgen, or thyroid hormone at this time.

3. *In vitro* assays cannot and will not be gatekeepers; they cannot constitute a decision node; they are useful as information for possible mechanisms (or site of action) but not as "yes/no" determinants to proceed to the *in vivo* screens or T2T because:
   a) *in vitro* assays mediated by receptor binding evaluate only one of many possible sites of action;
   b) negative results may mean relatively little due to limitations of the assays, e.g., lack of metabolic capability, solubility, etc. (i.e., false negatives); and
   c) positive results may be false positives.

4. Results from *in vivo* assays have more weight than results from *in vitro* assays since:
   a) *in vitro* assays will generate false negatives as well as false positives, based on differences in access to the target tissue, metabolism, etc. relative to *in vivo* assays; and
   b) *in vivo* results are considered to be more relevant.

5. Results from *in vitro* assays that assess endocrine EDC activity with and without metabolic activation are worth more than results from *in vitro* assays without metabolic activation (since the former can assess the activity of metabolites generated within the culture if the correct metabolic activation is used, e.g., rat liver S9, and the latter can only assay the parent compound).

6. Results from apical *in vivo* assays are worth more than results from specific *in vivo* assays (since they indirectly assay many more sites of action to get to the same endpoint; e.g., uterotrophic assay in ovariectomized adult females [specific assay; CSM must act at level of uterus] versus in intact immature females [apical assay; CSM can act at level of hypothalamus, pituitary, gonad, thyroid, and/or uterus]). A positive specific assay

provides mechanistic information (but other mechanisms of action may also be present and go undetected); a negative specific assay is less informative.

7. Biologically plausible results are worth more than biologically implausible results (obviously dependent on the state of our scientific knowledge).

8. A consistent pattern of positive (or negative) results in various related assays is worth more than a single isolated positive (or negative) result (e.g., positive results for binding to ER and transcriptional activation *in vitro* and positive results in an apical or specific uterotrophic assay *in vivo* are worth more than a positive result for receptor binding and transcriptional activation, but no uterotrophic response) (see additional comments in discussion of false negatives and false positives above).

9. The decision which will emerge from T1S is:
   a. The CSM does not require further testing is <u>not</u> a potential endocrine disruptor for E, A, or T activity at this time (the CSM goes to the "hold unless ..." box); or
   b. The CSM <u>should be tested further is</u> a potential endocrine disruptor for E, A, or T activity at this time, AND
      i. proceed to T2T; or
      ii. proceed to Hazard Assessment/Regulatory Action.

10. Statistical significance is a useful tool, but must be interpreted within the context of biological significance. For example, an observed association which does not achieve statistical significance, but which is consistent with results from related assays suggesting a common mechanism of action, might be interpreted as biologically significant. This means the use of any particular criterion such as P equal to 0.05 should be carefully considered, and there may be no hard and fast rule for weighting by statistical significance. Statistical significance (e.g., $p < 0.05$ in an appropriate, acceptable statistical test) is important and useful, but it must be interpreted within the context of biological significance (e.g., $p = 0.052$ in the assay, but consistent with results from related assays, some or all of which are statistically significant).

Insert Figure 5.1
Insert Figure 5.2
Insert Figure 5.3

V. Tier 2 Testing Concepts and Design Parameters

A. Introduction to T2T

The purpose of T2T is to characterize the nature, likelihood, and dose-response relationship of endocrine disruption of estrogen, androgen, and thyroid in humans and wildlife. T2T is a complement to T1S. As already discussed, T1S is composed of a battery of *in vitro* and *in vivo* assays designed to detect whether a substance may have EAT activity.the potential of a chemical to disrupt the endocrine system. The *in vitro* screening assays are highly sensitive

and quite selective for a particular mode of action. They are, however, quite far removed from the biological complexity of an intact animal and may give false positive readings have a tendency to give false positive readings because, for instance, not all substances which bind to a receptor will cause an adverse biological effect;, false negative readings may also result from the in vitro receptor binding or transcriptional assays because and not all endocrine disruptors act via a receptor. *In vivo* assays encompass the metabolic and response capability of a whole organism but focus on such a short time frame that the full effects of exposure to a chemical substance may not be identified and characterized. Since there is considerable biological conservation in the endocrine system, it is not necessary to screen in every major taxonomic group. Screens based on mammalian cell lines or intact animals will indicate the potential for adverse effects, which must be characterized in longer-term studies in species representing a variety of taxa.

T2T is the definitive phase of the screening and testing program and is intended to provide more detailed information regarding endocrine disruption activity of a tested CSM. Tests typically are less sensitive than screens which have been designed to value sensitivity more than specificityerr on the side of false positives as opposed to false negatives, thus one would expect more chemicals to test positive in the tier 1 assays than in the Tier 2 Tests. Primarily, this tier should assess the concentrations which elicit endocrine disruption and the consequences of such disruption to inform hazard risk assessments. To fulfill this purpose, tests are longer-term studies designed to encompass critical life stages and processes, a broad range of doses, and administration by a relevant route of exposure, so a more comprehensive profile of biological consequences of chemical exposure can be identified and related to the dose or exposure which caused them. Effects associated with endocrine disruption may be latent and not manifested until later in life or may not appear until the reproductive period is reached. Tests for endocrine disruption will encompass at least one generation including effects on fertility and mating, embryonic*in utero* or *in ovo* development, sensitive neonatal growth and development, and transformation from the juvenile life stage to sexual maturity. Unless a rationale exists to limit the mammalian test to one generation, mammals will be exposed to determine the effects of *in utero* and juvenile exposure to effects in a second generation.

Two-generation tests fully characterize potential effects of concern; however, there may be instances when less comprehensive study designs would be adequate (dependent on available prior information). Some of the considerations for determining whether the comprehensive two-generation tests or alternative tests would be conducted include an understanding of: the toxic profile of the chemical under study, mechanism(s) of action, exposure scenarios, use patterns, and populations at risk.

In selecting non-mammalian tests as with mammalian tests, EDSTAC chose those which would require the least modification from existing standardized methods. Thus the Tier 2 Tests for fish, birds, and amphibians are based on existing test guidelines. However, there are two differences: 1) more modification of the non-mammalian guidelines was required than for their mammalian counterparts and 2) the existing non-mammalian guidelines, in general, have not been widely and routinely run, as have the mammalian guidelines.

Existing guidelines for investigating developmental and reproductive endpoints are considered adequate for evaluating the most obvious consequences of endocrine disruption, but they will not pick up the more subtle effects. Unlike Tier 1 assays, Tier 2 tests do not address the specific mechanisms involved in toxicity, thus, developmental and reproductive toxicants operating through non-endocrine mechanisms may not be differentiable from those that do; however, this would be immaterial to a risk assessment per se.

B. Guidance for Selecting Tier 2 Tests

[NOTE TO THE READER: Concerns are still being raised as to whether the results from T2T will be definitive given the following guidelines. These guidelines make it possible to limit Tier 2 tests to a narrower set of species than would otherwise be the case, based on exposure related information. The concern raised is that while each of the individual tests are deemed definitive, if a chemical does not complete the full battery of tests (i.e., on all recommended species), it would not be possible to designate the chemical as either being an endocrine disruptor or not being an endocrine disruptor (for EAT) for the species that were not tested. These concerns are related to the definition of testing and the purpose of T2T (including its definitiveness) stated in ChapterThree, and the concerns regarding alternative approaches to T2T outlined in Section VI. C. of this chapter.]

The Conceptual Framework, found in Chapter Three, states that existing information on biological effects and exposure and the results of T1S should be used to inform decisions regarding the selection and design of Tier 2 tests. T1S information may be of use in determining whether to add a satellite assay for thyroid effects, for example, but may be of limited value in determining the selection of assays (e.g., bird reproduction and/or fish life cycle, etc.) since only a limited number of taxa are proposed for the *in vivo* assays and mammalian cell lines are used for *in vitro* assays in T1S. In this case, the choice of which Tier 2 tests to perform will consider the physico-chemical characteristics and environmental release and exposure information of the CSM to be tested, together with biological data from T1S.

As discussed in Chapter Three, EDSTAC has recommended that chemicals be allowed to voluntarily bypass T1S. These chemicals would, however, be required to complete the HTPS assays. Since T1S assays, in the aggregate, provide preliminary information on the presence and nature (mechanism) regarding endocrine activity the potential for endocrine disruption and the species and sex at risk, the absence of any such information as a result of bypassing T1S would mandate performance of all tests in the T2T battery (i.e., the mammalian and non-mammalian multi-generation tests with all the recommended endpoints). Performance of these tests would need to be consistent with the principles governing T2T which are set forth below.

In addition, it must be remembered that toxicity testing may be required, or even under way, for specific CSMs in order to satisfy legislative and regulatory requirements distinct from those included in the screening program mandated in the 1996 Food Quality Protection and

Safe Drinking Water Acts. As a consequence, implementation of testing designed to evaluate the potential for induction of effects related to interference with estrogen, androgen, and thyroid-mediated functions would be an overlay on existing requirements and may be subsumed in testing which is directed to the evaluation of a broader range of effects and modes/mechanisms of action.

As general guidance for which tests to perform, EDSTAC proposes the following principles:

1. Where use, exposure, and release of a substance are well characterized, it may be possible to tailor T2T for particular exposure scenarios. Conversely, if they are poorly characterized, positive results in T1S would trigger the entire battery of Tier 2 Tests unless other data clearly indicated that certain taxa would not be affected.

2. If sufficient data are available, an exposure assessment should be conducted to provide guidance in selecting Tier 2 Tests. Exposure assessments vary in scope and complexity, but "sufficient data" at a minimum includes chemical identity, basic chemical properties (water solubility, octanol/water partition coefficient, vapor pressure, Henry's Law constant), rates of significant transformation processes such as biodegradation, and use and release profiles. Measured values are preferable for chemical properties and environmental fate, but estimation methods are often satisfactory for supplying missing data. Ideally, the chemical use and release profile should provide information on the distribution of releases (if any) to air, water, soil, etc., and the amounts and frequency of such releases.

3. If a chemical is released to or can be predicted to reach streams, rivers, or lakes, a fish life cycle test with freshwater species and invertebrate life cycle test should be conducted. If release is to an estuary, marine species should be substituted for freshwater species in toxicity testing. If release is to both types of environment, freshwater species are preferred. The use and release profile obviously is essential in making this determination, but will also be valuable in determining which exposure models are used to make quantitative estimates of exposure.

   For example, freshwater aquatic exposure modeling often can be conducted using the Probabilistic Dilution Model (PDM) or a model such as REACHSCAN, which incorporates environmental degradation processes. PDM yields the frequency of exceedance of an ecological concern level preset by the modeler, and is useful not only when releases are from known point sources but also when only the *category* of use (via SIC codes) is known. At a higher level of complexity, even site-specific models such as EXAMS-II may be appropriate under some circumstances, for example, when there are only a few point sources and the site of release and downstream environments are well characterized.

   Soil fate models such as SESOIL and PRZM exist but also have extensive parameterization requirements. Short of this, rough estimates of mobility in soil and thus likelihood of reaching groundwater or surface waters that are in hydrological contact with

ground water can be made using certain screening-level tools routinely applied by EPA in, for example, PMN review.

For air releases numerous fate models such as ISCLT are available, and like the water models, they vary in complexity. They calculate concentrations of chemicals in air at assumed locations of human receptors (e.g., at certain distances from stack releases), but such exposure data may also be used as input for aquatic or terrestrial ecological exposure assessments.

Models for calculating environmental concentrations of chemicals released to estuaries are less well developed in general, but the model ESTUARY contains extensive hydrologic data for several major estuaries in the U.S. and is potentially useful. It was designed for high-volume consumer products that are widely used and dispersed in the environment (i.e., surfactants).

4. Pesticides with agricultural or other outdoor uses, and chemicals that would be expected to bioaccumulate and biomagnify through the food chain or that present a potential risk to birds, should be tested in the bird reproduction study.

[NOTE TO THE READER: At the plenary, the EDSTAC agreed to move the "Methods to Select the Target Doses for T2T" section (formerly "C") after the "Low Dose Considerations for T2T" section.

C. Low Dose Considerations for T2T

1. Introduction to the Issue

Serious issues have been raised as to the adequacy of classical approaches to regulatory toxicity testing that employ exaggerated dosing regimens (up to maximally tolerated dosages) in order to identify a hazard and extrapolate from these high doses to estimate risk or safe levels in the range of environmentally relevant exposures. There are two principle issues: 1) whether or not a threshold dose exists for receptor mediated toxicity, and 2) whether the dose response curve is monotonic or non-monotonic in nature.

With respect to the first issue, some believe that there is no threshold for effects of exogenous endocrine disruptors since there exist background levels of endogenous hormones that are already biologically active. Thus, any additional exposure constitutes an exceedance of the threshold. Evidence for this hypothesis includes findings of a study with turtle eggs in which very small amounts of additional estradiol shifted the sex ratio in a linear dose response manner (Ref). Breast cancer is also cited as an example in which endogenous levels of

estrogen are a known risk factor and it is argued that additional estrogenic substances would add to background levels of risk (Ref). Additional data to support the no-threshold hypothesis included the study of DES in mice which showed decreased fertility and number of pups with an increase in dose and similar studies with vinclozolin and TCDD in male rats (Ref).

The second issue relates to the shape of the dose-response curve. Although monotonic curves may vary in slope, the slope of such curves is always in the same direction (either positive or negative including zero) and therefore there are no local maximum or minimum points along the curve. With non-monotonic dose-response curves, local minima can exist such that an effect may be pronounced at low doses, then becomes statistically insignificant or disappears at intermediate doses, and finally reappears (or a different effect appears) at higher doses. The issue raised by non-monotonic curves is that "high dose" testing may fail to detect toxicity that occurs in the "low dose" region of the dose-response curve. This is because the classical approach to finding the NOAEL by progressive reduction of the dose beginning at high doses will locate the nadir (i.e., a local minimum at some intermediate dose) but will not locate the second region of increased effect at doses below the apparent NOAEL. This issue is further complicated by the possibility of different effects at low doses as compared to high doses. There are examples of this phenomenon for endogenous hormones. For example, it is well known that testosterone stimulates sperm production up to a point, but at excess levels inhibits it (Elwell, ref.) There is evidence that the developing mouse prostate responds in a non-monotonic manner to estrogens, in that prostate weight is increased initially, then decreased by higher maternal dosages of potent estrogens like DES and ethinyl estradiol. There is one report in the literature using a limited number of mice indicating that the weak estrogen bisphenol A has effects similar to these potent estrogens at low doses, but not at higher doses (Nagel et al, 1997).

The example that fat-soluble vitamins and essential trace elements exhibit non-monotonic dose-response relationships has been used as support for the widespread existence of such relationships. However, the situation with these agents is fundamentally different than the examples cited above. For the hormones cited above, the adverse effects observed, whether increases or decreases from baseline, all occur at hormone levels above those necessary for normal function. This is not the case for the vitamins and essential minerals. For those compounds the non-monotonicity occurs at either end of the normal range; i.e., the nadir in the dose-response curve for adverse effects does not represent a false NOAEL, but is the range of
normal nutrition.

Examples of this phenomenon are the actions of certain endogenous substances, including vitamins, trace elements, and hormones, which exhibit non-monotonic dose-response curves (Ref). By analogy to the action of these endogenous substances, it is argued that exogenous substances (commercial chemicals, pesticides, and contaminants), particularly those that can

mimic a vitamin or hormone, may also exhibit non-monotonic dose responses. Some recent studies on the effects on prostate weight and sperm production in male offspring of female mice treated with estrogen, DES, and bisphenol A display an inverted U-shaped dose response and are cited as evidence for this phenomenon (Ref).

There is intense scientific debate surrounding these issues that centers on two principal questions: First, are data implicating xenobiotics in such phenomena reproducible and broadly generalizable to endocrine endpoints and endocrine active chemicals, and second, is the low dose phenomenon indicative of adverse effects at the individual or population level? If low dose phenomena are reproducible, generalizable, and related to adverse effects, the implications for regulatory toxicity testing and risk assessment are profound. It should be recognized that there are divergent scientific opinions on the "low dose" issue at the present time and that more research is necessary to answer these questions.

The EDSTAC notes that, historically, testing has sometimes missed critical endpoints either by 1) failing to dose during the most sensitive life stage (Morrissey, et al, 1987); 2) failing to test in a susceptible organism (Chamberlin 1979, Fraser 1988); or 3) failing to examine subtle (yet biologically important) endpoints. For example, early studies on the developmental effects of PCBs in rodents identified fetotoxicity as a critical endpoint, yet these studies failed to test at low enough doses or to measure subtle enough endpoints. As a result these early studies missed the neurotoxic effects of PCBs which occur at much lower doses.(Tilson et al. 1990).

These omissions in testing may lead either to missing a critical effect completely, or to identifying a NOAEL which is excessively high. The EDSTAC has attempted to minimize the likelihood of these types of errors by requiring testing in a variety of organisms during sensitive life stages. A variety of endpoints which appear to be low-dose sensitive have also been added to the EDSTP testing protocols. In addition, a number of endocrine disrupter-sensitive endpoints have already been added to EPA's reproductive toxicity testing guidelines that will be issued this spring.

The Committee agrees, however, that dose selection in T2T must include special attention to setting the low dose. In particular, the low dose should not be selected by identifying the high dose and then dropping the dose by a fixed formula of a couple of orders of magnitude. Instead, a number of considerations need to go into selecting the low dose, including the results of prior information, including HTPS, toxicity testing, pharmacokinetic, and epidemiology data, where available. Information about environmental exposure levels might also be used where appropriate. Finally, range-finding studies should be constructed so as to identify both the high dose and the low dose for testing, with inclusion of low-dose sensitive endpoints in range-finding. These precautions will minimize the likelihood that critical effects will be missed or that excessively high NOAELs will be identified in T2T.

2. Recommended Research Program

EDSTAC recommends that additional research to resolve existing controversies about the nature of the dose-response curve for endocrine active substances, particularly with regard to the low dose region, be given high priority. Several endocrine active substances should be selected to systematically evaluate:

a) Selected potential effects on males (e.g., prostate weight, sperm counts, etc.) and females (e.g., ovarian follicular development) exposed *in utero* and/or via lactation over a wide range of doses;
b) Whether or not effects (if any) persist throughout the entire lifetime of the species tested, and the long-term significance of any effects observed (i.e., are they adverse to the long-term health of the animals or populations?);
c) Nature of the dose-response curve for any effects observed, with a particular focus on the low dose region (i.e., do very low doses give greater responses than the NOAEL andor is there no observable threshold?); and
d) Species and sex comparisons among proposed T2T species/tests.between rats and mice.

EDSTAC recommends that a collaborative program involving government, industry, and appropriate individuals in academia design the study protocols, be kept abreast of the conduct of the studies, evaluate results, and develop overall conclusions and recommendations.

[NOTE TO THE READER (from EDSTAC members based on an assignment given at the March plenary): A more detailed description of additional research needs, including studies designed to replicate recent studies on low-dose effects, is likely to emerge from the May 11-13 workshop, "Characterizing the Effects of Endocrine Disruptors on Human Health at Environmental Exposure Levels." This workshop is organized by the National Institute for Environmental Health Sciences in cooperation with others. The output from this workshop, and from ancillary activities including current testing and research will be pertinent to the low dose issue when this information becomes available., will be considered by EDSTAC for possible inclusion and endorsement in its June 1998 final report. Timely funding and execution of additional research could provide results that would influence initial implementation of T2T.]

[NOTE FROM EPA STAFF: The EDSTAC should try to make the above list as definitive as possible. However, we may want to defer this issue until our June meeting at which time we may want to endorse the research recommendations of the May 11-13, 1998, expert workshop "Characterizing the Effects of Endocrine Disruptors on Human Health at Environmental Exposure Levels." If so, the list of topics included above may need to be modified based on the recommendations from the workshop.]

3. Interim Measures

[NOTE TO THE READER: In considering whether to recommend an interim measures program the EDSTAC has discussed two alternative approaches for identifying chemicals

that would be subjected to such a program, however further discussion is necessary to resolve the issue. For this reason, the following language has been bracketed to indicate the unresolved nature of these discussions.]

[The research program outlined above will take several years to complete and should be conducted in parallel with the validation and standardization of the T1S battery and the implementation of phase 1 of screening. If we assume it will take two years for validation and standardization of the screening battery and one additional year to complete screening, the first chemicals that go through screening will not be ready for T2T before 2001. Although questions raised about potential low dose effects of endocrine disruptors and questions regarding appropriate dose levels in T2T may be resolved by that time, an interim policy is needed for: 1) those chemicals which will bypass T1S and proceed to T2T in the meanwhile and 2) for other substances in case the research program takes longer to complete than the standardization and validation of the T1S battery. Delay in resolving these issues has potential consequences for chemicals going through testing and assessment; they may have regulatory or advisory levels set too high to be protective of public health and the environment, and it could be necessary that they be tested again at lower doses.

The EDSTAC agreed that the EPA should not incorporate additional lower doses and low dose sensitive endpoints for routine regulatory testing until further research provides answers to the questions posed above. Nevertheless, EDSTAC agrees on the following points which it believes can form the basis of an interim approach:

a) The concern for effects at ultra-low doses and non-monotonic response curves only involves toxicity that is mediated by receptors. All of our experience in toxicology indicates that the potent toxicants act through receptors (endocrine and non-endocrine). The literature which is cited as supporting this concern for endocrine disruptors involves the estrogen receptor specifically.

b) The high throughput pre-screen can be used to identify those chemicals which may present a low-dose concern. These assays have been specifically designed to detect binding to the receptor and in the case of the transcriptional activation assays, a consequence (agonism or antagonism) of that binding. The assays are exquisitely sensitive to low doses of natural hormones and hormone-mimics.

c) **Low dose effects have not been observed in the absence of a related toxic effect observed at traditional high doses.** Compounds that appear to show estrogenic effects at low doses are estrogenic at high doses although the effects may be different.

The EDSTAC is recommending the following interim approach for setting doses in T2T pending resolution of the low dose issue. Substances being considered for bypassing T1S, which have a potent positive response in one or more of the HTPS assays would be subjected to a wide range of doses, including multiple doses in the low-dose region, in range-finding studies for T2T and would be a high priority for T2T.

Substances slated to go through T1S that test positive in HTPS assays with high potency will be designated as high priority for full T1S Screening. Based on the "weight-of-evidence" criteria from the HTPS assays together with the full T1S battery, these substances with high potency would move to become high priority for T2T. Range-finding studies for T2T would include endpoints for which there is evidence of low-dose sensitivity from existing literaturemultiple doses in the low dose region. If effects are seen at low doses in the range-finding studies, multiple doses in the low-dose range would be included in T2T.

If research confirms the general applicability of low dose phenomena and links them to adverse effects, EDSTAC would expect EPA to modify its testing and hazard assessment policies accordingly.]

D. Methods to Select the Target Doses for T2T

For T2T in mammals, other vertebrates, and invertebrates, the EDSTAC proposes that the methods to select doses used in the performance of these tests include:

previous information such as that available during the priority setting phase including results from the HTPS (or its equivalent by bench-level assays);
results from T1S (including the range-finding study results);
results from other assays or tests for pesticide registration, etc.; and
results from range-finding studies.

Suggested uses for 1, 2, and 3 are presented in Section III. A. Range-finding studies specifically for T2T should be performed at multiple doses (at least five) with a limited number of animals per dose, an abbreviated duration (which must include exposures during gestation or egg development and lactation), and a limited number of relevant endpoints including low dose sensitive ones. If further research validates the low dose concern, EDSTAC would recommend the inclusion of low dose sensitive endpoints for the range-finding study to determine the need for inclusion of low doses in the definitive T2T. Endpoints, identified in recent publications, which, at present, appear to be low dose sensitive include: prostate weight (for mammals), epididymal sperm concentration (for mammals), other accessory sex organ weights (all vertebrates), thyroid weight (all vertebrates), reproductive capability (all T2T), and vaginal threads (for mammals). All of these, except for vaginal threads, are included in the 1996 guidelines; however, vaginal threads would be identified during the examination of offspring females for vaginal patency. New and/or different low dose sensitive endpoints may be identified as new data are generated.

Current toxicological test guidelines generally require testing at a minimum of three dose levels plus a control. These guidelines specify that the top dose level should be a maximally tolerated dose (MTD), that is, a dose which by definition is toxic but which does not result in excessive mortality (not to exceed 10%). In reproductive and developmental toxicity studies, the MTD is usually based on parental or maternal toxicity, which is expressed as depressed body weight gain, actual weight loss, reduced feed and/or water consumption,

treatment-related clinical signs of toxicity, etc. The MTD is set based on available toxicity information such as data from a range finding study. The next lower dose is ideally set at an intermediate toxic dose and the lowest dose at a level at which no toxic effects are observed. If additional lower doses are included in the study because of identified concerns at low doses, the additional doses should be widely spaced (perhaps an order of magnitude) to identify the nature of the dose-response curve in the low dose region. The toxicity upon which the MTD is based may or may not be related to the endpoints, which are the object of the investigation (e.g., cancer, neurotoxicity, reproductive, or developmental effects).

The results of range-finding studies mustshould be discussed with the Agency prior to performance of the definitive studies and/or included in the submission of the T2T results for evaluation by the Agency.

E. Testing Antithyroid Activities in T2T

Thyroid hormones are well known to play essential roles in vertebrate development (Dussault and Ruel, 1987; Myant, 1971; Porterfield and Hendrich, 1993; Porterfield and Stein, 1994; Timiras and Nzekwe, 1989). Experimental work focused on the effects of thyroid hormone on brain development in the neonatal rat supports the concept of a "critical period," during which thyroid hormone must be present to avoid irreversible damage (Timiras and Nzekwe, 1989). Though the duration of this critical period may be different for different thyroid hormone effects, the general view has developed that this is the period of maximal developmental sensitivity to thyroid hormone, and it occurs during the lactational period in the rat (Oppenheimer et al., 1994; Timiras and Nzekwe, 1989). Although thyroid hormone receptors are expressed in fetal rat brains (Bradley et al., 1989; Strait et al., 1990), and thyroid hormone can exert effects on the fetal brain (Escobar et al., 1990; Escobar et al., 1987; Escobar et al., 1988; Porterfield, 1994; Porterfield and Hendrich, 1992; Porterfield and Hendrich, 1993; Porterfield and Stein, 1994), the lactational period represents a stage of rapid expansion of the central nervous system that coincides with a large increase in the expression of thyroid hormone receptors (Bernal et al., 1985), and an increase in the number of demonstrated effects of thyroid hormone on brain development.

In conducting thyroid-related tests in tier 2 the EDSTAC recommends using an approachFor the reasons outlined above, the EDSTAC recommends testing potential thyroid disruptors in a paradigm in which dosing occurs during the fetal and lactational period. In addition, there are a variety of endpoints that would provide reliable markers of thyroid disruption. Brain weight offers a simple measure, though it is not thyroid specific. Characteristics of myelination, or of myelin basic protein expression (either mRNA or protein), would provide a more selective measure (Bhat et al., 1981; Bhat et al., 1979; Farsetti et al., 1991; Figueiredo et al., 1993; Rodriguez-Pena et al., 1993; Shanker et al., 1987). In this regard, the expression of myelin basic protein and/or neurogranin/RC3 may offer the simplest and most specific endpoints of thyroid disruption during the perinatal period (Farsetti et al., 1991; Iniguez et al., 1993). These mRNAs are both enormously abundant and robustly affected by thyroid hormone. However, their sensitivity to xenobiotics has not been studied. A list of existing

endpoints for thyroid hormone function, and additional ones recommended by EDSTAC for validation and inclusion, are found in Table 5.3 on p. 5-53.

VI. Proposed Tier 2 Testing Battery

A. Outline of Proposed T2T Battery

The T2T battery includes the two-generation reproductive toxicity study or a less comprehensive test and tests addressing at least four taxonomic groups, including birds, amphibians, fish, and invertebrates.

Mammalian Tests

Two-Generation Mammalian Reproductive Toxicity Study; or
A Less Comprehensive Test:
Alternative Mammalian Reproductive Test; or
One-Generation Test.

Tests for Other Animal Taxa

Avian Reproduction (with bobwhite quail and mallard)
Fish Life Cycle (fathead minnow)
Mysid Life Cycle (Americamysis)
Amphibian Development and Reproduction (Xenopus)

B. Two-Generation Mammalian Reproductive Toxicity Study

The two-generation reproductive toxicity study in rats (TSCA 799.9380 [August 15, 1997]; OPPTS 870.3800 [Public Draft, February 1996]; OECD no. 416 [1983]; FIFRA Subdivision F Guidelines - 83-4) is designed to comprehensively evaluate the effects of a chemical on gonadal function, estrous cycles, mating behavior, fertilization, implantation, pregnancy, parturition, lactation, weaning, and the offspring' ability to achieve adulthood and successfully reproduce, through two generations, one litter per generation. Administration is usually oral (dosed feed, dosed water, or gavage), but other routes are acceptable with justification (e.g., inhalation). In addition, the study also provides information about neonatal survival, growth, development, and preliminary data on possible teratogenesis. The experimental design for a two-generation reproductive toxicity study is presented in Figure P.1, which is found in Appendix P, Tier 2 Test Study Designs.

In the existing two-generation reproductive toxicity test, a minimum of three treatment levels and a concurrent control group are required. At least 20 males and sufficient females to produce 20 pregnant females must be used in each group as prescribed in this current guideline. The highest dose must induce toxicity but not to exceed 10% mortality. In this study, potential hormonal effects can be detected through behavioral changes, ability to

become pregnant, duration of gestation, signs of difficult or prolonged parturition, apparent sex ratio (as ascertained by anogenital distances) of the offspring, feminization or masculinization of offspring, number of pups, stillbirths, gross pathology and histopathology of the vagina, uterus, ovaries, testis, epididymis, seminal vesicles, prostate, and any other identified target organs. Table 5.3 provides a summary of the endpoints evaluated within the framework of the experimental design of the updated two-generation reproductive toxicity test (and some recommended additional endpoints for validation and inclusion s, still under consideration, to cover EAT concerns).

These observations are comprehensive and cover every phase of reproduction and development. Tests that measure only a single dimension or component of hormonal activity, (e.g., *in vitro* or short-term assays) provide supplementary and/or mechanistic information, but cannot provide the breadth of information listed in Table 5.3, which is critical for risk assessment.

Additionally, in this study type, hormonally induced effects such as abortion, resorption, or premature delivery as well as abnormalities and anomalies such as masculinization of the female offspring or feminization of male offspring, can be detected. Substances such as the phytoestrogen, coumesterol, and the antiandrogen cyproterone acetate, which possess the potential to alter normal sexual differentiation, were similarly detected in this study test system (i.e., 1982 Guideline). The initial prebreed exposure period (10 weeks) of the two-generation reproductive toxicity test also provides information on subchronic exposures which can be used for other regulatory purposes.

Table 5.3

Mammalian Two-Generation Endpoints

Below are two types of lists: first, those of the endpoints required in current EPA test guidelines (1996); and second, additional endpoints recommended by EDSTAC for validation and inclusion, which will detect estrogen, androgen, and thyroid hormone perturbations. EDSTAC recommends these additional endpoints be incorporated into the two-generation mammalian reproduction test guidelines when validated.

Circumstances under which all of the endpoints would be evaluated include: 1) the CSM sponsor has chosen to bypass T1S and 2) prior information and/or the results of T1S indicate the potential for effects on thyroid hormone function.

Current Guideline Endpoints Sensitive to Estrogens/Antiestrogens

sexual differentiation
gonad development (size, morphology, weight) > accessory sex organ (ASO) development
ASO weight ± fluid; histology
sexual development and maturation: vaginal patency (VP), preputial separation (PPS)
fertility
fecundity

time to mating
mating and sexual behavior
ovulation
estrous cyclicity
gestation length
abortion
premature delivery
dystocia
spermatogenesis
epididymal sperm numbers and morphology; testicular spermatid head counts; daily sperm production (DSP); efficiency of DSP
gross and histopathology of reproductive tissues
anomalies of the genital tract
viability of the conceptus and offspring (maintenance of implantation)

Recommended Additional Endpoints for Validation and Inclusion

accessory sex organ function (secretory products)
sexual development and maturation (nipple development and retention)
androgen and estrogen levels
LH and FSH levels
testis descent (?)

Current Guideline Endpoints Sensitive to Androgens/Antiandrogens

altered apparent sex ratio
malformations of the urogenital system
altered sexual behavior
changes in testis and accessory sex organ weights
effects on sperm numbers morphology, etc.
nipple retention in males (or retained nipples in male offspring)
altered AGD (now triggered from PPS/VP)
reproductive development; PPS/VP (puberty)
male fertility
agenesis of prostate
changes in androgen-dependent tissues in pups and adults (not limited to sex accessory glands)

Current Guideline Endpoints Sensitive to Thyroid Hormone Agonists/Antagonists (general)

growth, body weight
food consumption, food efficiency
developmental abnormalities
perinatal mortality

testis size and DSP
VP; PPS

Recommended Additional Endpoints for Validation and Inclusion

neurobehavioral deficits (see developmental landmarks below)
TSH, T4, thyroid weight and histology (e.g., goiter)
developmental landmarks:

- prewean includes pinna detachment, surface righting reflex, eye opening, acquisition of auditory startle, negative geotaxis, mid-air righting reflex, motor activity on PND 13, 21, etc.
- postwean includes motor activity PND 21 and postpuberty ages (sex difference); learning and memory PND 60 - active avoidance/water maze
- brain weight (absolute), whole and cerebellum
- brain histology

C. Alternative Approaches to Mammalian T2T

[NOTE TO THE READER: While the EDSTAC has discussed the concept of alternative approaches to mammalian T2T in some detail, a number of concerns still exist with the specifics in this section. Some changes were made by a small group at the end of the last plenary meeting, however these changes do not address all of the remaining concerns. Two questions in particular have not yet been addressed: 1) how does the information obtained from the two-generation test differ from that obtained in one or the other alternative tests? and 2) assuming you have an answer to the first question, under what criteria should be used to determine which chemicals can be tested in one or the other of the alternative tests?

It has been suggested that the standardization and validation program should help answer the first question. If the EDSTAC agrees, this should probably be made explicit in the text. Regarding the second question, the EDSTAC still needs to have further discussion. The second paragraph below begins to raise the criteria; however, some members would like to see more transparent language. In particular, concerns have been raised with the use of production volumes, exposure criteria, and/or resource limitations as reasons to run the alternative tests. In addition, some members feel further refinement of this section should be coordinated with the final review and refinement of the definition of testing and purpose of T2T stated in Chapter Three, and the guidelines to selecting Tier 2 Tests contained in Section V. B. of this chapter. Finally, concerns with the inclusion of references to the EU approach still exist.

EPA staff recognize some of the difficulties in defining, absolutely, the range of instances when a chemical could move through the "alternative" tests. Within the next few weeks, they intend to develop and disseminate their thoughts on how to resolve the issues included in this section. In addition, it may be necessary to pull together a small group of EDSTAC members to try to develop an agreement to present to the rest of the plenary prior to the June meeting.]

The standard two-generation reproductive toxicity test fully characterizes potential effects of concern. The mammalian multigeneration test proposed for Tier 2 is the most comprehensive test available for assessing reproductive and developmental toxicity. It is already mandated for some pesticides, food additives, and, in Europe, very high production new chemicals (>1,000 tons per annum; tpa). The estimated cost for running the test according to the new harmonized TSCA-FIFRA guidelines is $350,000 to $800,000 per chemical, depending on route of administration. This money and resource allocation is certainly warranted for biologically active materials, substances intended to be taken internally and chronically, and for extremely high production volume/potential for exposure chemicals. However, bBecause of resource limitations, it should not be the only test available or we run the risk of evaluating fewer chemicals for reproductive and developmental toxicity potential. This is contrary to our goal of public health protection.

There are instances when a less comprehensive study design would provide comparable information, when considered with existing data, on which to make a decision. Some considerations for determining whether the comprehensive two-generation reproductive toxicity test or an alternative test would be conducted include an understanding of: mechanisms of action; exposure scenarios; use patterns; populations at risk; and other prior information. Examples of circumstances under which one of the alternative tests might be run include: 1) a full two-generation reproductive toxicity study has been run in the past, but it either was conducted in accordance with the "old" guideline and/or the results in the previous study require additional follow-up that is best accomplished using one of the alternative protocols; 2) production volume and potential for exposure is low; or 3) there is low probability of at-risk populations being exposed. In cases 2 and 3, the alternative test may be used more as a preliminary evaluation and, therefore, not necessarily as the last evaluation of the potential for EAT effects in a reproductive toxicity study. Therefore, EDSTAC has included alternative, less comprehensive study designs below.

In running these alternative tests, the need to protect public health and the environment must be recognized. The cost of a single test or series of laboratory assays (screens or tests) is considered in light of the cost, over time, to human health and environmental impairment from exposure to toxic chemicals and mixtures.

There are precedents for conducting reproductive and developmental toxicity testing in such a sequential manner. The OECD's Screening Information Data Set (SIDS) program for international high production volume chemicals (>1,000 tpa) includes a one-generation screen for reproductive and developmental endpoints. If effects are observed in the screen, the OECD considers production volume, and exposure potential in OECD countries to determine the priority for confirmatory testing for either or both endpoints. The European Union (EU), for example, employs a framework for testing of new chemicals in which more and more comprehensive tests are conducted as production volume and potential for exposure increase. This framework is overlaid with considerations, based on expert judgment, that can increase the pace with which a chemical moves through the testing framework. Chemicals for which the potential for widespread exposure is great, or that bear structural similarities to known hazards, move through the system more rapidly. For example, a one-generation test must be

conducted on all new chemicals that reach a production volume of 100 tpa, but may be required with production volumes of only 10 tpa if the level of concern, based on exposure or toxicity, is high. Furthermore, a compound that produces adverse or equivocal results in a one-generation or developmental toxicity test has a high priority for additional, more comprehensive testing. Thus, tTonnage production triggers are often seen as useful ways of requiring toxicological screens or tests; however, triggers will not account for toxicity. Innate toxicity must be taken into account when establishing testing triggers based on production amount.

Under TSCA, testing can be required for new and existing chemicals based on high production volume and exposure (reference 4a1b Policy Statement) without consideration of hazards and risk. Thus, the two-generation study can be required for chemicals presenting high volume and exposure. In cases where the chemical presents both hazard issues and high volume and exposure issues, the test can be required at relatively lower levels of production. Thus, while Tthere are no explicitly required data sets for new chemicals submitted under TSCA, but decisions on the extent of data required are made using similar decision criteria as the EU (i.e., other toxicity data, concern for the toxicity of the chemical class, production volume, potential for exposure and widespread distribution, persistence, etc.). The main practical difference is the absence of explicit bright line standards imposed by tonnage triggers.

Two Proposed Alternative Tier 2 Tests:

EDSTAC acknowledges that the developing organism may be uniquely sensitive to the effects of endocrine-active agents. Therefore, any mammalian Tier 2 Test should include a careful assessment of the consequences of *in utero* and lactational exposure on subsequent growth and development.

Although developmental toxicological endpoints are not specifically required for assessment of EAT toxicity, both of the alternative, less comprehensive mammalian tests can be modified to assess these endpoints where appropriate based on previous information such as T1S results, overall production, use and exposure scenarios, and/or to meet other regulatory requirements for a given CSM.

For either of these alternative designs, assessment can be performed by terminating a portion of the F0 dams in each group just prior to anticipated parturition (i.e., on gd 20-21) and performing gestational and fetal structural evaluations (i.e., ovarian corpora lutea, uterine implantation sites, total, resorbed, dead and live fetuses, live fetal number, sex, weight, external, visceral, and skeletal alterations), i.e., a “standard” developmental toxicity evaluation by OPPTS 1996 draft guidelines; USEPA TSCA guidelines 870.3700, 1997; FDA guidelines, 1993.

1. Alternative Mammalian Reproduction Test

A graphical representation of the study design (Figure P.2), as well as additional text, for the Alternative Mammalian Reproduction Test (AMRT) is provided in Appendix P, Tier 2 Test Study Designs. The objectives of this test are to describe the consequences of *in utero* and/or lactational exposure on reproduction and development from compounds that displayed EAT activity in the T1S. It is intended that this test could replace the standard EPA multigenerational reproductive test (TSCA guidelines, 1997) in T2T. In this regard it will be conducted with at least three treatment groups plus a control and includes endpoints sensitive to chemicals that alter development via EAT activities.

The AMRT involves exposure of maternal rats (designated F0 generation) from gestational day 6 (time of implantation), through partuition (birth), and through the lactation period until weaning of offspring (designated F1 generation) on postnatal day 21. F1 offspring (both sexes) are retained after weaning with no exposures for 10 weeks and then mated within groups. F1 males are necropsied after the mating. F1 females and their litters (designated the F2 generation) are retained until the F2 generation is weaned. F0 females (and a subset of F1 weanlings) are necropsied with organ weights and possible histopathology. F1 animals are evaluated for reproductive development (VP, PPS), estrous cyclicity, and, at necropsy, for organ weights, possible histopathology, andrological assessments, and T3/T4 (with TSH triggered). F2 weanlings are counted, sexed, weighed, examined externally, and discarded.

This alternative test differs from the "standard" two-generation study design in that this test:

does not include exposures prior to mating, during mating, or during the early preimplantation stage of pregnancy in the dams;
does not include exposures to parental males; and
does not include direct exposure to the postweanling offspring; potential exposure is limited to *in utero* transplacental and/or lactational routes.

This alternative test differs from the other alternative in that this study design provides for:

exposure to the F0 dam only from gd7 (sperm detection = gd1), through weaning of the F1 offspring on pnd 21;
no exposure to parental males;
mating of the F1 animal (who have not been directly exposed) to produce F2 offspring; and following the F2 offspring to weaning on pnd 21.

2. One-Generation Test

A second alternative to the standard two-generation reproductive toxicity test is a one-generation reproductive toxicity test. A graphical representation of the one-generation test (Figure P.3), and additional text, is provided in Appendix P, Tier 2 Test Study Designs. It has been used in rats and mice; most labs have experience in one or more of the following rat

strains: CDâ (Sprague-Dawley), Fischer-344, Wistar, and Long-Evans. It has been, and is currently being, used as a range-finding study prior to performance of a guideline two- (or more) generation study for the last 10 years under EPA (TSCA/FIFRA) GLPs; design similar to that used by Sharpe et al. (1996). This is a shortened, scaled-own version of the new draft OPPTS and Final TSCA guidelines for reproductive toxicity testing.

The one-generation test less comprehensively evaluates *in utero* development, but has the advantage of assessing adult reproductive capacity and postnatal development. The postnatal leg of the test is extended to evaluate vaginal patency and preputial separation in offspring. However, retention of Mullerian duct derivatives may be difficult to assess in this protocol (and in the two-generation assay, for that matter).

The one-generation test involves a short prebreed exposure period for male and female rats of the initial parental generation (designated F0), exposure continuing through mating, gestation, and lactation of F1 litters. F0 males are necropsied after F1 deliveries; F0 females are necropsied after F1 weaning. Postweanling F1 animals are directly exposed for a 10-week postwean period and are then necropsied. F1 animals are evaluated for reproductive development (VP, PPS), estrous cyclicity and at necropsy for organ weights, possible histopathology, andrological assessments, and T3/T4 (TSH triggered). F0 animals will undergo the same necropsy assessments.

This alternative test differs from the "standard" two-generation study design in that this test:

does include a prebreed exposure period, but it is shorter (basic design calls for two weeks, can be prolonged) than in the standard two-generation study (10 weeks to encompass one full spermatogenic cycle in rats); and
does include direct exposure of F1 offspring after weaning, including exposure through puberty and sexual maturation, but does not evaluate effects of in utero and/or lactational exposure (and beyond) on generation of F2 offspring. F1 male and female reproductive organs (weight/histology), estrous cyclicity, and andrological endpoints are assessed at scheduled necropsy on PND 90 ±2.

This alternative test differs from the other alternative in that this study design provides for:

exposure to both male and female F0 parental animals prior to mating, during mating, and during gestation and lactation of F1 offspring (F0 males are necropsied after F1 deliveries, F0 females are necropsied after F1 weaning);
direct exposure of postweanling F1 offspring after lactation until termination; and
no mating of F1 animals to produce F2 offspring.

D. Description of the Tests for Other Animal Taxa

The EDSTAC agrees T2T should address at least four other animal taxonomic groups, including birds, amphibians, fish, and invertebrates. It is recommended that the following

standardized tests be used as a basis for a non-mammalian battery:

Avian reproduction (with bobwhite quail and mallard)
Fish life cycle (fathead minnow)
Mysid life cycle (Americamysis)
Amphibian development and reproduction (Xenopus)

Except for the amphibian study, these tests are routinely performed for chemicals with widespread outdoor exposures and expected to affect reproduction. Modifications to each may be warranted to enhance the ability to detect endocrine-related effects. The amphibian test, though not standardized, is considered warranted because of the extensive fundamental knowledge base on amphibian development and reproduction.

Just as for mammalian testing, there may be instances when less comprehensive study designs would be adequate. Considerations for determining whether the full battery of comprehensive mammalian tests include an understanding of mechanisms of action, environmental fate and transport, persistence, potential for bioaccumulation, and potential ecosystems exposed.

Production volume is also a consideration for less comprehensive approaches. Comprehensive assessments of environmental toxicity, including chronic toxicity assays in a variety of species, are already generated for pesticides and very high production volume chemicals (>1,000 tons per annum) in Europe. The EU explicitly requires less comprehensive assessments for lower production volume chemicals, with additional testing required as production increases. As with mammalian assessments, these moving triggers recognize that potential for exposure is correlated with production volume. While there are no explicitly required data sets in the U.S. under TSCA, similar decisions are made on data adequacy based on other information, including production volume.

There are a number of alternative, less comprehensive assays that may be appropriate to consider for environmental toxicity assessments. These might include shorter-term avian development tests (like the one that is discussed later in this document as a research need), and the fish early life cycle test, and Daphnia reproduction test, both of which are already established protocols.

1. Avian Reproduction Test

While birds are not included as subjects in the T1S battery, it is important to evaluate the effects of exposure of birds to CSMs with endocrine activity. Furthermore, birds are fundamentally different from mammals in the control of sexual differentiation (males are the homogametic sex) so results using mammalian subjects will not provide complete information relevant to birds.

Use of the EPAs Avian Reproduction Test guidelines (OPPTS 850.2300) is recommended, modified to include the additional endpoints presented below to make the test more sensitive

to CSMs with endocrine activity. Table 5.4 provides a summary of the endpoints evaluated within the framework of the Avian Reproduction Test (and recommended additional endpoints for validation and inclusion to cover EAT concerns). Two important extensions of this guideline are recommended: 1) modification and standardization of the husbandry and dosing of the offspring from EPAs Avian Reproduction Test guidelines (OPPTS 850.2300) to create a two-generation avian reproduction test; and 2) using the procedures of the modified Avian Reproduction Test protocol, evaluate an additional exposure pathway (i.e., direct topical exposure, which is common in the wild, by dipping eggs). The recommended extensions to the guideline are outlined in Appendix P.

In the current Avian Reproduction Test guidelines, two species are commonly used, mallards and northern bobwhite. Exposure of adults begins prior to the onset of maturation and egg laying and continues through the egg-laying period; their offspring are exposed, in early development, by material deposited into the egg yolk by the females. These offspring can be used efficiently to test for the effects of CSMs on avian development. There are several endpoints currently required [see OPPTS 850.2300, (c) (2)] that are particularly relevant to disruption of endocrine activity, including: eggs laid, cracked eggs, eggshell thickness, viable embryos, and chicks surviving to 14 days. The guidelines should be extended with additional observations made for circulating steroid titers, thyroid hormones, major organ (including brain) weights, gland weights, bone development, leg and wing bone lengths, ratios of organ weights to bone measurements, skeletal x-ray, histopathology, functional tests, and reproductive capability of offspring (Baxter, et al, 1969; Bellabarba et al. 1988; Dahlgren and Linder, 1971; Emlen, 1963; Cruickhank and Sim, 1986; Fleming et al., 1985a; Fleming et al., 1985b; Fox, 1976; Fox et al., 1978; Freeman and Vince, 1974; Hoffman and Eastin, 1981; Hoffman and Albers, 1984; Hoffman, 1990; Hoffman, et al., 1993; Hoffman, et al., 1996; Jefferies and Parslow, 1976; Maguire and Williams, 1987; Martin, 1990; Martin and Solomon, 1991; McArthur et al., 1983; McNabb, 1988; Moccia, et al., 1986; Kubiak et al,. 1989; Rattner et al,. 1982; Rattner et al,. 1987; Summer et al,. 1996; Tori and Mayer, 1981). Other avian assays were considered including the Japanese quail androgenic assay (proctodeal gland), egg injection, draft OECD Japanese quail reproduction, and two generation avian reproduction tests, but were not selected because the endpoints addressed were limited or there was a lack of accepted and standardized methods.

Table 5.4

Avian Reproduction Test Endpoints

Current Guideline Endpoints Sensitive to Estrogens/Antiestrogens, Androgens/ Antiandrogens, and/or Hypothalamic-Pituitary-Gonadal Axis

egg production
eggs cracked
viable embryos (fertility)
eggshell thickness
fertilization success
live 18-day embryos
hatchability

14-day-old survivors

Recommended Additional Endpoints for Validation and InclusionProposed Additions
sex ratio
major organ (including brain) weights
gland weights
bone development (skeletal x-ray)
ratio of organ weights to bone measurements
histopathology
plasma steroid concentrations
neurobehavioral test (cliff test)
cold stress test
nest attentiveness

Current Guideline Endpoints Sensitive to Thyroid Hormone Agonists/Antagonists

body weight of adults
food consumption of adults
body weight of 14-day-old survivors
developmental abnormalities

Recommended Additional Endpoints for Validation and InclusionProposed Additions
plasma T3/T4
thyroid histology
bone development (skeletal x-ray)
ratio of organ weights to bone measurements
neurobehavioral test (cliff test)
cold stress test

2. Fish Life Cycle Test

The freshwater fathead minnow *Pimephales promelas* is the recommended species to be used and is continuously exposed from fertilization through development, maturation, and reproduction, and early development of offspring with a test duration of up to 300 days. The fathead minnow is also the recommended species for use in the screening battery for the fish gonadal recrudescence assay, and as such, the relevance of any activity detected in the screening assay would be evaluated. However, EDSTAC proposes a performance-based approach to species selection and, as more appropriate species are developed and validated, EDSTAC strongly encourages their use. For example, if exposure to a particular CSM is predominantly estuarine or marine, the estuarine sheepshead minnow *Cyprinodon variegates* may be substituted since experience and an established method exist for this species.

Fish are the most diverse and least homologous to mammals of all vertebrates. Reproductive strategies extend from oviparity, to ovoviviparity, to true viviparity. The consequences of an

endocrine disruptor may be quite different across the many families of fishes. As a first step though, only a fathead minnow, or in special cases the sheepshead minnow, life cycle test is suggested to confirm and quantify any effects detected by the Tier 1 battery. Subsequent tests with other species will then be a function of the risk assessment and nature of the hormones involved and effects expected/obtained.

The fish life cycle test (OPPTS 850.1500)follows procedures outlined in Benoit (1981) for the fathead minnow and Hansen et al. (1978) for the sheepshead minnow. In general, the test begins with 200 embryos distributed among eight incubation cups in each treatment group. When hatching is completed, the number of larvae are reduced to 25 individuals, if available, which are released to each of four replicate larval growth chambers. Four weeks following their release into the larval growth chambers, the number of juvenile fish are reduced again and 25 individuals, if available, distributed to each of two replicate adult test chambers. When fish reach sexual maturity, fish are separated into spawning groups (pairs or one male/two females) with a minimum of eight breeding females. Remaining adults will be maintained in the tank but will be segregated from the spawning groups. Adults will be allowed to reproduce, at will, until the 300th day of exposure. Alternatively, the test may be continued past 300 days until one week passes in which no eggs from any group have been laid. The embryos and fish are exposed to a geometric series of at least five test concentrations, a negative (dilution water) control, and, if necessary, a solvent control.

Assessment of effects on offspring of the parental group (first filial or F1 generation) will be made by collecting two groups of 50 embryos from each experimental group and incubating those embryos. When embryos hatch, the number of larvae hatched from each group will be impartially reduced to 25, if available, and released into the larval growth chambers. After four weeks of exposure, lengths and weights of surviving individuals will be made.

Observations are made of the effects of the test substance on embryo hatching success, larvae-juvenile-adult survival, growth of parental and F_1 generation, and reproduction of the adults. Table 5.5 provides a summary of the endpoints evaluated within the framework of the Fish Life Cycle Test (and recommended additional endpoints for validation and inclusion to cover EAT concerns).

Table 5.5

Fish Life Cycle Test Endpoints

Current Guideline Endpoints Sensitive to Estrogens/Antiestrogens, Androgens/Antiandrogens, and/or Hypothalamic-Pituitary-Gonadal Axis

viability of embryos
time to hatch
spawning frequency
egg production
fertilization success

Recommended Additional Endpoints for Validation and InclusionAdditions recommended for research and development
sexual differentiation (tubercle formation, gonadal histology)
sex ratio
gonadosomatic index
gamete maturation (production, final oocyte maturation, sperm motility test, etc.)
vitellogenin
plasma steroid concentrations
in vitro gonadal steroidogenesis

Current Guideline Endpoints Sensitive to Thyroid Hormone Agonists/Antagonists

growth, length, and body weight
developmental abnormalities

Recommended Additional Endpoints for Validation and InclusionAdditions recommended for research and development
plasma T3/T4
neurobehavioral deficits
thyroid histology

3. Mysid Life Cycle Test

Invertebrates constitute the majority of the fauna, but the relevancy of EAT actions to these organisms and the availability of tests to evaluate such is limited. However, it is plausible that a CSM which interferes with estrogen or androgen actions could interfere with ecdysteroid activity which is an important steroid in arthropods. The mysid life cycle test which has been standardized (OPPTS 850.1350) ; ASTM) E####) would allow a determination of the relevancy of an EAT active material to the development, molting, growth, and sexual reproduction in this important group of invertebrates. The saltwater mysid *Americamysis bahia* is tentatively preferred to the freshwater daphnid *Daphnia magna* because this species undergoes a full sexual reproductive cycle where the daphnid is parthenogenic in the standardized assay. The daphnid reproduction test (OECD 202) is well standardized, widely used for evaluating conventional toxicity of general and pesticidal chemicals, is less resource intensive than the mysid test, and daphnids have been demonstrated to respond to estrogenic compounds (Baldwin et al. 1995; Baldwin et al., 1997; Shurin and Dodson, 1997). However, the lack of sexual reproduction reduces the comprehensive utility of the daphnid. Other invertebrates, such as molluscs and echinoderms, do have EA systems, but again relevant standardized tests for evaluating the consequences of interfering with these systems are not currently available.

4. Amphibian Development and Reproduction

A definitive test with an amphibian, which exposes larvae through metamorphosis and reproduction, is important to evaluate the consequences of endocrine disruption in a poikilothermic oviparous vertebrate distinct from fishes. A rich literature on metamorphosis, growth, and reproduction exists for frogs and promising methods are being developed. No established method has been identified which is suitably comprehensive to stand as a T2T, and as such, this test falls in the Category V status of assays. The EDSTAC feels a test to address this taxonomic group and set of endpoints is needed in T2T and should be given a high priority for development and standardization.

VII. Summarizing the Interconnections Between HTPS, Bypassing T1S, Low Dose Concerns, and the Definitiveness of T2T

A. Context and Time Period Within Which These Issues Ultimately Will be Resolved

The complex nature of many of the issues discussed in this Report can make the connections between numerous issues fairly difficult to follow. The EDSTAC believes one particularly intricate thread throughout the document is the interconnection between the recommendation to perform HTPS, the scenarios by which some chemicals may bypass T1S, the approach to addressing low dose concerns, and the definitiveness of T2T. This section is intended to briefly reintroduce the major components of the thread and succinctly lay out the resulting implications.

The EDSTAC believes the interconnections between these issues are most acute for those chemicals that would go through T2T during "Phase I" of the EDSTP. As described in Chapter Seven, "Phase I" of the EDSTP will commence after completion of the standardization and validation of both T1S and T2T. EPA estimates it will take a minimum of two years, and perhaps as long as three years, to complete the standardization and validation process. By definition, the only chemicals that will go through T2T during Phase I are those permitted to bypass T1S. Furthermore, it is estimated it may take as long as one to two years for any one chemical to complete T2T. The EDSTAC feels these are important practical realities that should be kept in mind in reading this section.

B. High Throughput Pre-Screening

EDSTAC is recommending that all chemicals currently produced in quantities over 10,000 lb. per year and all substances that will bypass T1S (i.e., will go straight to T2T) go through the HTPS. As explained in Chapter Four, Section V. the HTPS has several potential uses:

as a component of T1S to identify CSMs that interact with EAT receptors;
as a source of biological effects information that can be used to assist in priority setting;

as a source of biological effects information that can be used to improve quantitative structure activity relationship (QSAR) models, thereby potentially reducing the extent and cost of future screening; and

d) as a means of identifying chemicals that interact with hormone receptors at low doses.; and

e) to provide information on the low dose region in range-finding studies for T2T.

The last two items are particularly important in the design of Tier 2 Tests and form the link between HTPS and the other issues discussed in this section.

C. Alternative Means of Obtaining T1S Information Versus Bypassing T1S

As noted in Chapter Three, the EDSTAC expects the vast majority of chemicals included in the EDSTP will go through the program in the logical, hierarchical manner for which the program was designed. Notwithstanding this expectation, the EDSTAC recognizes there will be circumstances where it may be inefficient to follow all steps of the EDSTP.

EDSTAC believes the EDSTP should be sufficiently flexible to allow a test sponsor to bypass screening assays. The EDSTAC has identified four circumstances, introduced in Chapter Three, Section X. B, where a chemical substance may not be required to perform the assays included in the recommended T1S battery. The EDSTAC believes it is helpful to distinguish the first two scenarios, which are in essence alternative means of meeting the information requirements associated with T1S, from the latter two scenarios, which are considered "bypassing T1S." Each scenario is discussed below:

1. Existing Information is Sufficient to Move to Hazard Assessment

The EDSTAC recommends that chemicals already subjected to tests that are the "functional equivalent" of the two-generation tests, ED endpoints, taxa, and dosing considerations recommended by the EDSTAC for T2T, be allowed to bypass both T1S and T2T and go directly to hazard assessment.

2. Alternative Means to Meet T1S Information Requirements

The EDSTAC also recommends it be permissible to complete the information requirements of T1S through the submission of data that are "functionally equivalent" to the data that would be generated from the recommended T1S battery. Further, functionally equivalent

information could be submitted for one or more of the recommended T1S assays or for the entire battery.

[NOTE TO THE READER: This text was pulled from the section previously entitled "Skipping T1S." It was recommended that the language be included somewhere in this section and the facilitation team worked with EDSTAC members to determine where it best fit.]

The concern that the screening assays and tests may not be of equal sensitivity should not be of concern so long as the tests are conducted at the proper dose levels (see Section IV. C. on dose considerations). T1S assays provide information on possible mechanisms of EDC activity (and therefore areas of concern for T2T), on relative potency (and therefore initial doses to be considered for T2T), on possible target organisms and environments at risk (and therefore which T2T to perform if T2T is tailored), on potential low dose concerns if known endpoints sensitive to low dose concerns are affected (and therefore whether to include low doses in T2T). If this information is available from other sources, then T1S (in whole or in part) does not have to be performed. If this information is not available from other sources, components of the T1S will have to be performed.

3. Bypassing T1S

There are two scenarios in which the EDSTAC recommends the owner of a chemical should be permitted to bypass T1S. Each of these scenarios has different implications for the information requirements associated with completing T2T and hazard assessment following T2T.

Chemicals Previously Subjected to Two-Generation Toxicology Tests

This scenario (which is discussed in Section II. C. of Chapter Four), includes food-use pesticides tested prior to implementation of the 1996 Toxicity Testing Guidelines, and may also include a small number of industrial chemicals. The EDSTAC recognizes the 1983 FIFRA Testing Guidelines, under which much of the information for these compounds was obtained, do not sufficiently capture estrogen, androgen, and thyroid-sensitive endpoints now recommended for T2T. EDSTAC believes the owners of chemicals meeting this criterion should develop a proposal regarding the additional studies (which could include a combination of screening-level assays, alternative T2 testing protocols, and/or other tests, as appropriate) to be conducted on the chemical to augment the previously conducted two-generation test(s).

The EDSTAC has recommended that chemicals meeting this criterion still be subjected to the HTPS assays, for the reasons discussed above. Chemicals meeting this criterion will also be

the most likely candidates for the alternative approaches for completing T2T, as discussed in Section VI. C. of this chapter.

b) Chemicals for Which There is No Prior Toxicology Testing

This scenario includes those chemicals where the owner of the chemical has decided to voluntarily achieve the definitive results of T2T without having completed the full T1S battery or any prior two-generation toxicology testing. By definition, under this scenario chemicals will have the least amount of information available prior to conducting T2T, since they will not have passed through T1S, nor will they have functionally equivalent T1S or T2T data.

EDSTAC has recommended that these chemicals also be required to complete the HTPS assays. Since T1S assays, in the aggregate, provide preliminary information on the presence and nature (mechanism) regarding endocrine activitythe potential for endocrine disruption and the species and sex at risk, the absence of any such information as a result of bypassing T1S under this scenario would mandate the broadest coverage in T2T (i.e., performance of the two-generation mammalian test with all endpoints and the tests for other taxa). Performance of these tests would need to be consistent with the principles governing T2T which are set forth in Section V. B. of this chapter.

D. Low Dose Concerns

As noted in the previous section, serious issues have been raised as to the adequacy of classical approaches to regulatory toxicity testing that employ exaggerated dosing regimens (up to maximally tolerated dosages) in order to identify a hazard and extrapolate from these high doses to estimate risk or safe levels in the range of environmentally relevant exposures. These issues are controversial and cannot be answered without further research.

The HTPS bears directly on both the low dose testing issue and the bypass issue because it will be the only mechanism by which potent activity at low doses is identified for substances that bypass T1S. It will also be the means of determining the low doses to be administered in range-finding studies. Bypassing T1S and the low dose issue are linked further because some substances that bypass T1S may be identified as a concern at low doses before the research is completed and must be dealt with in some interim fashion. EDSTAC's recommendations for an interim policy are discussed in Section VII. C. 2. of this chapter.

E. Definitiveness of T2T and the Interconnections Between the Issues

Questions and concerns have been raised during EDSTAC deliberations about the definitive nature of T2T results. These questions arise in connection with the determination of a "weight-of-evidence" conclusion for T2T in the absence of T1S results (i.e., under one of the bypass scenarios) or in the presence of "equivocal" T2T results in combination with a "positive" T1S result. EDSTAC believes it is critically important to design the screening and

testing program such that there will be a definitive result coming out of T2T. In particular, it should be designed so that a negative result coming out of T2T will override any prior positive results in HTPS and/or T1S.

Many screening assays are mechanistically specific and are thus capable of identifying chemicals that may act by a specific mechanism or mode of action. They are designed to be especially sensitive since the goal is to minimize false negatives. Tier 2 Tests, on the other hand, are more apical and less sensitive, but reflect the real-world complexity of toxicity in whole organisms (e.g., absorption, distribution, metabolism, excretion) and response at the level of the organism (e.g., growth, development, behavior). In the "weight-of-evidence" approach, results of *in vivo* assays outweigh those of *in vitro* assays and, in the EDSTAC parlance, tests outweigh screens. Properly conducted Tier 2 Tests are intended to be the final arbiter of whether or not a substance is an endocrine disruptor; that is, when the results of T2T are unambiguous, they provide the definitive answer. The expression of adverse effects is a necessary condition at the level of T2T is a necessary condition for designation as an endocrine disruptor. In some cases, it will be apparent that the type of effect seen in T2T is a result of endocrine disruption; in other cases, the results of T2T will not allow us to make that judgment because it could be the result of a different mode of toxicity. T2T focuses on identifying agents that act as reproductive or developmental toxicants through the endocrine-mediated mechanisms considered by EDSTAC. However, it may not always be possible to identify the mechanism by which a specific manifestation of toxicity is produced in T2T. Such information can be determined through further research using more focused assays, some of which may be drawn from T1S. In bypassing T1S we are losing mechanistic information that could tell us whether adverse effects seen in T2T are the result of endocrine disruption or not; however, labeling a substance an endocrine disruptor is not essential for EPA or other regulatory agencies to conduct a hazard assessment and take risk management action.

As stated in the Chapter Three, Section IX. C., where there is a definitive negative result from T2T, a chemical will be placed in the "hold box." In such a circumstance no further screening and testing would be required for the chemical unless:

a) existing statutes require periodic review (e.g., FIFRA re-registration);
b) new statutory requirements mandate review;
c) new screens or tests for endocrine disruption are incorporated into the EDSTAC strategy, which will generate significant new information, or invalidate prior screens or tests upon which decisions have been made to stop screening and testing; and/or
d) new information on the endocrine disrupting activitypotential of the chemical substance or mixture becomes available and it is determined that this new information warrants additional testing.

It is inevitable with toxicological testing that equivocal results will sometimes be obtained in T2T. When this occurs one must look at the possible reasons for the ambiguity and see if there are ways to resolve it. Occasionally, conducting other tests, or running assays to investigate the mechanism of action (if T1S was bypassed), will resolve the ambiguity. Alternatively, repeating the study, perhaps at different exposure levels, will resolve it. In

other cases, one of the abbreviated tests may provide the information necessary to allow a more informed "weight-of-evidence" determination of the hazard potential of the CSM.

In summary, EDSTAC agrees that substances canshould be permitted to bypass T1S under certain circumstances because they will be evaluated in the battery of tested in definitive Tier 2 tests. These tests that will include a wide range of apical and specific endpoints to detect for known adverse effects from EAT disruption in a number of representative species across vertebrate and invertebrate taxa. The new Tier 2 test guidelines being recommended by EDSTAC will identify the CSMs provide assurance that chemicals with adverse effects from EAT disruption will be identified, within the limits of based on the current state of the science.

VIII. Standardization, Validation, Methods Development, and Research

A. Concept of Assay Validation and Standardization

As stated earlier, the role of standardization and validation is to provide sufficient data to allow informed decisions about the relative merits of the proposed T1S battery component assays and alternative assays (based on sensitivity, specificity, technical complexity, inter- and intra-laboratory variability, time, and cost).

Validation is the scientific process by which the reliability and relevance of an assay method are evaluated for the purpose of supporting a specific use (ICCVAM, p. 15). Relevance refers to the ability of the assay to measure the biological effect of interest. Measures of relevance can include sensitivity (the ability to detect positive effects), specificity (the ability to give negative results for chemicals that do not cause the effect of interest), statistically derived correlation coefficients, and determination of the mechanism of the assay response with the toxic effects of interest. Reliability is an objective measure of a method's intra- and inter-laboratory reproducibility. The process of validation includes standardization, that is, definition of conditions under which the assay is run (species, strain, culture medium, dosing regimen, etc.). Standardization is critical to ensure reliability, that is, valid results from time to time and between laboratories. Even in those instances where there is currently some degree of *de facto* acceptance of a given screening method as valid, there is a need for such standardization.

B. Statutory Need for Validation

The Food Quality Protection Act (FQPA) requires EPA "to develop a screening program, using appropriate validated test systems and other scientifically relevant information, to determine whether certain substances may have an effect in humans that is similar to an effect produced by a naturally occurring estrogen, or other such endocrine effect as the Administrator shall designate" by August 1998 with implementation of the peer review

program by August 1999. In requiring the use of validated test systems the FQPA is merely mandating good science. There are numerous reasons for using only validated assays. These include: having confidence that the assay is detecting the effect it purports to be detecting, that the results of the assay are reproducible and comparable from laboratory to laboratory, and that the results permit a comparison of the toxicity of various chemicals. These factors are important in being able to interpret results to establish a relative priority among chemicals for progressing from screening to testing and ultimately to perform a hazard and risk assessment.

C. Addressing the Validation Issue

The assays being considered by the EDSTAC vary considerably in terms of their degree of development and validation. EPA (Dr. Lynn Goldman, April 24, 1997, letter to EDSTAC) recognized that few screening assays have actually met the "gold standard" of validation and that other assays have been accepted on the basis of peer review and general use without formal validation. Because the number of validated off-the-shelf assays is so limited, EDSTAC was asked by EPA to extend its consideration to all existing assays.

Thus, although formal validation is not a prerequisite for assay selection by EDSTAC, the ability or potential of an assay to be validated must be considered because assays must be capable of passing the "validation test" before the screening program is fully implemented. The following are a list of factors that EPA might consider in estimating the likelihood that a candidate assay will actually survive the validation process. If possible, it would be useful to devise a quantitative or semi-quantitative scale for expressing these characteristics so that test methods reviewed by various people could more easily be compared.

Number of independent, peer-reviewed publications reporting results from the assay.
Similarity of results from independent publications performing the assay.
Number of independent laboratories publishing results from the assay.
Consistency of the methods used across laboratories.
Consistency of results of the assay between labs (to the extent results are available for the same chemicals).
Known variability of the assay within single laboratories (may not preclude use as a valuable research tool, but may have important implications for use as a widespread screening tool).
Age of the assay --is it an "old standby" or a "new kid on the block."
The extent to which the assay relies upon calibrated equipment and calibrated standards?
The extent to which the assay depends upon the skill of the technician (a professional opinion from those who know about the assay).
The extent to which the assay utilizes internal controls or standards.
Use of the assay to develop clinically effective drugs (although not of itself proof of validity, success speaks well for itself).
Number of examples of false positives/false negatives from the assay (although we have a clear charge to eliminate false negatives at the screening stage, it is nonetheless

important to consider the overall performance of an assay in order to estimate its likelihood of being validatable).

Any known species-or strain-specific sensitivities of the assay. For *in vitro* assays, any knowledge of critical sensitivity to cell characteristics such as passage number, plating density, doubling time, etc. or other specific sensitivities of the assay, such as receptor number, transfection technique, serum requirements, media composition, etc. (trying to get another angle on how finicky the assay is).

Information related to validation status of the assays is summarized in Section VIII. E. and was among the factors considered in deciding among assays. These same considerations regarding validation and standardization apply to T2T.

D. Validation and Standardization Process

While not all assays would necessarily need to be validated since they may have *de facto* acceptance as valid in the scientific community due to their long history of use and performance, others have little or no data that would allow judgments to be made regarding their validity.

The following is a description of elements of a such a validation process for endocrine screening assays.

1. Characterize Reference Substances/Vehicles.

A reference substance for each hormone endpoint (minimum)
Reference substances for each hormone endpoint:
positive control substances
negative control substances
Natural and man-made substances
Composition and purity defined (e.g., GC/MS)
Stability verified
Coded
Centralized distribution (using same batch number, lot number, etc.)

2. Develop a Standard Protocol for Each Assay Method.

Assay system to be used (species, strain, sex, age, cell line, clone, gene construct, etc.)
Dose levels/exposure concentrations
General criteria for selection
MTD (whole animal systems; e.g., mortality, decreased body weight, etc.)
Cell viability (*in vitro* systems): 2 methods
Solubility limitations
Specific for each reference substance
Dose/exposure regimen

Number of doses
Duration
Guidance of mixtures
Concentrations of reference substances
Interpretation of data
Route of exposure (whole animal systems)
Description of endpoint(s) to be measured
Materials (equipment, media, vehicles, etc.)
Time(s) of measurements
Criteria for positive/negative response
Statistical methods to used
Number of replications required (depending on the study design)

3. Define Specialized Skills and Equipment Required for Each Assay Method

4. Conduct in a Variety of Laboratories.

5. Compile and evaluate data

Expert scientific oversight group
Inter-laboratory/intra-laboratory variability
Positive versus negative response
Relative potency (in comparison to reference substance)
Viability/maintenance of assay system (including passage number and growth curves)
Sensitivity of assay system (e.g., minimal effective dose/concentration)
Specificity of assay system (positive versus negative controls)

E. Current Validation Status of Screens and Tests

As mentioned throughout the chapter, each assay and test under consideration in T1S or T2T needs some level of standardization, validation, methods development, or further research before being accepted as a regulatory toxicity screen or test. The level of standardization and validation varies according to a variety of criteria applied to each of the assays, including: period of time in use, existing level of general acceptance in the endocrine toxicology field, and existing understanding of relevancy and reliability. The assays proposed for T1S and T2T fall into five general categories with regard to the level of standardization, validation, or methods development required.

[NOTE TO THE READER: Questions have arisen regarding the level of standardization and validation necessary for the mammalian, avian, and fish tests when the recommended additional endpoints are considered. Efforts will be made to resolve these differences before

the next draft of the report is released.]

Category I: Screens and tests which have been fully validated and standardized are placed in Category I. These procedures meet all the criteria of relevance and reliability for use in regulatory toxicity screening or testing for estrogen, androgen, and thyroid. As such, these screens and tests are recommended by EDSTAC as the initial components of EPA's EDSTP. As other procedures become sufficiently standardized and validated to warrant inclusion in Category I, EDSTAC recommends that such screens and tests be incorporated into the EDSTP according to their specific and appropriate use. Currently, only the following tests are included in Category I:

Two-Generation Mammalian Reproductive Toxicity Study (1996 Public Draft Guidelines and 1997 TSCA Final Guidelines)
One-Generation Test

Category II: Screens and tests which have been in use for a sufficient period of time and which have gained sufficient general acceptance within the field of endocrine toxicology to be considered *de facto* validated (reliable AND relevant) are included in Category II. These assays measure relevant endpoints, are responsive to endocrine active compounds with a high degree of specificity, are sufficiently sensitive to identify all known active agents, and can reasonably be expected to give reproducible results from laboratory to laboratory, assuming a general level of competence and expertise. Nonetheless, variations in protocols for these screens and tests can produce disparate results. Therefore, standardization of the protocol to be recommended for these screens and tests should be accomplished by EPA before these assays are implemented as screening requirements for endocrine activity or disruption. Currently, the following screens and tests are included in Category II:

ER Binding Assay
AR Binding Assay
Rodent 3-Day Uterotrophic Assay (Subcutaneous)
Rodent 5-7 Day Hershberger Assay
Rodent 3-Day Uterotrophic Assay (Intraperitoneal);
Avian Reproduction Test (with Bobwhite Quail and Mallard) (as currently performed)
Fish Life Cycle Test (Fathead Minnow) Test (as currently performed)
Mysid Life Cycle Test (Americamysis)

Category III: Screens and tests which have sufficiently broad use to be generally considered relevant OR reliable to either screening for endocrine activity (Tier 1) or to testing for adverse endocrine-mediated effects (Tier 2) are included in Category III. These assays cannot, however, be generally considered to be both relevant AND reliable. The level of performance that can be expected of these assays with respect to identifying endocrine active agents or endocrine disruptive effects of chemicals must be clarified. Therefore, these assays should undergo further but focused validation and standardization to define their relevance and reliability for the task of endocrine disruptor screening or testing. The validation required may be focused to answer specific questions about relevance and to provide information regarding

specificity and sensitivity. Currently, the following screens and tests are included in Category III:

ER Transcriptional Activation Assay
AR Transcriptional Activation Assay
Steroidogenesis Assay with Minced Testis
Rodent 20-Day Pubertal Female Assay With Thyroid
Placental Aromatase Assay
Rodent 14-day Intact Adult Male Assay with Thyroid
Rodent 20-day Thyroid/Pubertal Male Assay
Alternative Mammalian Reproduction Test
Avian Reproduction Test - when performed in multiple generations
Turtle Egg Assay

Category IV: Screens and tests which may have relevance to the task of either screening for endocrine activity or testing for endocrine disruptive effects, but whose performance in identifying endocrine active agents or endocrine disruptive effects has seen only limiting testing are included in Category IV. Questions as to whether these assays measure endpoints that are relevant to endocrine activity or endocrine disruptive effects, whether these assays respond with specificity and sensitivity to known endocrine active agents, or whether they identify endocrine disruptive effects cannot be addressed with information currently available. In addition, questions regarding the specific protocols and conditions under which the assays should be conducted must be answered before relevance and reliability can be assessed. Nonetheless, the EDSTAC feels that these assays would have sufficient utility, if further developed and validated, to enhance or augment the screening and testing program. Therefore, the EDSTAC recommends that resources be made available to pursue methods development and validation and standardization of these assays. Currently, the following screens and tests are included in Category IV

Frog Metamorphosis Assay
Fish Gonadal Recrudescence Assay
14-Day (PND 9-22) Developmental/Thyroid Assay

Category V: Screens and tests which, if available, could have an important utility in the screening and testing program are included in Category V. However, such assays have not actually been conducted. Therefore, the EDSTAC recommends that research be conducted to determine whether such assays can be developed, and if so, what purpose the assays could fulfill within the endocrine disruptor screening and testing program. Rationale for including each of these assays and tests is found below. Although the EDSTAC recognizes additional research priorities may become important in the future, the following were identified as research priorities:

in utero developmental screening assay
in ovo developmental screening assay

avian androgenicity screening assay
invertebrate screening assays
amphibian development and reproduction test (Xenopus)
reptilian reproduction test

In Utero Developmental Screening Assay: The EDSTAC recognizes the importance of evaluating postnatal consequences of *in utero* exposures to EDCs with EAT activities. Such an assay has not been incorporated into T1S due to the lack of an appropriate, short-term, cost-effective assay. Research is needed to design, standardize, and validate such an assay. One such study design is as follows:

F0 rat females are exposed by gavage on gd 7 (sperm detection = gd 1) through pnd 20. At parturition (PND 0) document number of total (live and dead) pups, apparent sex ratio, and anogenital distance is measured in F1 offspring (increased in females from androgen, decreased in males from estrogen or antiandrogen); one female pup per litter is necropsied on PND 0 and 5 for uterine weight (uterine weight will be increased with exposure to an estrogenic compound). Litter sizes are standardized to eight pups with as equal a sex ratio as possible on PND 5 after second female per litter is necropsied. On PND 10-12, male pups are examined for retained nipples (males will exhibit retained nipples with exposure to an antiandrogenic compound). On PND 14, one female pup per litter will be necropsied with uterine weight and gland number; increased uterine weight, decreased gland number, and increased luminal epithelial cell height will result from exposure to E2-agonists (e.g., Ki 67 or PCNA can be used to evaluate cell cycle). On PND 21, all remaining F1 pups will be terminated; male pups will be examined for testicular ectopia, hypospadias, and other reproductive tract anomalies (due to anti-androgenic compounds and possibly estrogens); female pups will be examined for precocious puberty (e.g., vaginal patency), uterine and ovarian weight, and urogenital anomalies (from estrogens or androgens). For both sexes, blood samples will be taken at PND 21 necropsy for T4, TSH (thyroid/antithyroid), and estradiol in females and testosterone in males. Myelin basic protein (MBP) will be measured by dot-blot or Northern blot in F1 pup brains after *in utero* exposure (if the MBP assay is verified, standardized, and validated it may be a useful addition to mammalian tier 2 tests which involve *in utero* exposure to identify thyroid effects). Throughout lactation, periodic body weights will be recorded for pups and dams on PND 0, 5, 10, 15, 20, and 21. Maternal animals will be necropsied on PND 21 with uterine implantation sites counted, and thyroid weighed and histology; maternal blood samples will be taken for T4 and TSH (detect thyroid/anti-thyroid activity).

Duration of study:
in life:
1 week quarantine
1-2 weeks mating
3 weeks gestation
3 weeks lactation
8-9 weeks total

Minimum 10 dams/group; suggest five groups; estimated cost $50,000.

In Ovo Developmental Screening Assay: A major route of excretion of lipophilic contaminants for female birds is into the yolk of their eggs, their avian embryos can have high levels of exposure from the earliest stages of development. In addition, the endocrine control of sexual and reproductive development is fundamentally different in birds than in mammals. Hence, a short-term screening assay for CSMs that alter avian development is highly desirable. There is a moderate amount of research on the effects of environmental contaminants injected into bird eggs that could be the basis for developing such an assay.

Avian Androgenicity Screening Assay: This assay would be useful in the T1S battery to improve and extend our assessment of CSMs for androgenic and antiandrogenic activity in birds. Development of this screen will become important if data from T2T point to differences in the actions of CSMs in birds versus mammals.

Reptilian Reproduction Test: Several distinctive features of reptilian reproduction (e.g., ovoviviparity, temperature-dependent sex determination), and a generally long life span that allows high body burdens of environmental contaminants to accumulate in reptiles, underscore the importance of developing a practical reproductive test in this class of ecologically important vertebrates.

F. Instituting a Validation Program

EDSTAC recommends that a multi-stakeholder process involving government, industry, and academics be utilized to standardize and validate the T1S and T2T batteries. One key step in instituting a validation program for T1S assays is the identification of a set of "standard test substances" for the individual assays as well as for the overall T1S battery. To the extent possible, the standard test substances will be chosen according to the following criteria:

- known EAT positives which act via receptor binding
- known EAT positives that do not appear to act via receptor binding (i.e., via some other mechanism such as alterations of hormone synthesis, degradation, transport, etc.)
- known EAT negatives (i.e., substances known not to have hormonal activity)
- known EAT positives which are active as the parent compound
- known EAT positives which require metabolic activation
- substances that cover a wide range of EAT potencies
- substances with a wide range of physical properties (pH, reactivity, volatility, etc.)
- substances with extensive *in vivo* databases with *in vivo* effects that have been well documented

It may not be possible to satisfy every one of the above criteria (e.g., there are currently no known examples of environmental thyroid or androgen receptor agonists), but every standard test substance selected should meet at least one of the criteria.

In addition, careful definition of the expected use of the set of chemicals is necessary to avoid inappropriate use. Such a set of chemicals, developed with the already mentioned criteria in mind, would be used in the validation program to assist in defining their relevance and reliability for the task of endocrine disruptor screening, i.e., to identify whether a specific CSM is a potential endocrine activitydisruptor, or can be placed in the "hold unless ..." box.

Further, as was also stated earlier, it is critical to acknowledge state of the science in this area is evolving rapidly, and assays currently being developed, or ones developed in the future, may offer distinct advantages over some included in the current options. As they are developed, validated, and standardized, the use of these new assays for screening is strongly encouraged.