DRAFT version 5.4

July 16, 2007

Validation of Screening and Testing Assays Proposed for the EDSP

I. Introduction

A. Issue

The United States Environmental Protection Agency (U.S. EPA) is responsible for administering Federal statutes for protecting both human health and the environment. In the case of chemical substances, the U.S. EPA employs various physical-chemical, biological, and toxicological assays or methods to generate information in order to assess a chemical substance. It is essential that these assays and methods provide the U.S. EPA with reliable and correct information in order that the U.S. EPA fulfills its responsibilities to protect both human health and the environment. The ability of the methods and assays to provide correct information on a chemical substance in a reliable and consistent manner is demonstrated and assessed through a process called validation. This paper will describe the Endocrine Disruptor Screening Program (EDSP) and how commonly accepted validation criteria will be interpreted and applied to the assays intended for use by the U.S. EPA in the EDSP.

B. The Basis of the EDSP

Section 408(p) of the Federal Food Drug and Cosmetic Act (FFDCA) 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 may designate [21 U.S.C. 346a(p)].

Upon recommendations from the Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC), the EDSP was expanded using the Administrator's discretionary authority to include the androgen and thyroid hormone systems and wildlife effects. In accepting the EDSTAC's recommendations (63 FR 71542; December 28, 1998), EPA adopted a two-tiered screening program. The purpose of Tier I is to identify the potential of chemicals to interact with the estrogen, androgen, or thyroid hormone systems. The purpose of Tier II is to identify and characterize the adverse effects resulting from that interaction and the exposures required to produce them.

EDSTAC recommended a number of assays for EPA consideration as potential Tier I screens and Tier II tests for detecting and characterizing endocrine disrupting chemicals. The Committee recognized that a Tier I battery should have the following characteristics:

- The Tier I screening battery should maximize sensitivity to minimize false negatives while permitting an as of yet undetermined, but acceptable, level of false positives.

- The Tier I 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 substances or mixtures are not overlooked.
- The Tier I 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, and thyroid hormones should be detected.
- The Tier I battery should include a sufficient range of taxonomic groups among the test organisms. There are known difference in endogenous ligands, receptors and response elements among taxa that may affect endocrine activity of chemical substances or mixtures.
- The Tier I battery should incorporate sufficient diversity 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.

EDSTAC's recommendation for a Tier I screening battery included the following in vitro and in vivo assays:

Estrogen receptor binding or transcriptional activation
Androgen receptor binding or transcriptional activation
In vitro steroidogenesis assay

Uterotrophic
Hershberger
Pubertal female with thyroid
Frog metamorphosis assay for thyroid
Fish screening assay

EDSTAC recognized there were other combinations of assays that might substitute for some components of the recommended battery and also recommended that EPA validate the following assays as alternatives:

Placental aromatase assay
Intact adult male assay with thyroid
Pubertal male assay with thyroid

For Tier II tests EDSTAC recommended the following:

One- or two-generation mammalian reproductive toxicity test
Avian reproduction
Fish life cycle
Amphibian development and reproduction
Mysid (invertebrate) life cycle

EDSTAC stated that the guiding principle for the treatment of false positives and negatives should be to value sensitivity more than specificity (more tolerance of false positives than false negatives) at the screening level and consider Tier II to provide the needed specificity (distinguish between true and false negatives). EDSTAC noted that false results may arise from the stochastic nature of screens and tests and the limitations of assays.

The EDSP is described in detail on the following website: http://www.epa.gov/scipoly/oscpendo/

C. Requirement for Validation

As noted above, section 408(p) of the FFDCA requires EPA to use validated test systems. Validation has been defined as "the process by which the reliability and relevance of a test method are evaluated for a particular use" (OECD, 1996; NIEHS, 1997).

> *Reliability* is defined as the reproducibility of results from an assay within and between laboratories.
> *Relevance* describes whether a test is meaningful and useful for a particular purpose (OECD, 1996). For the EDSP Tier I, relevance is defined as the ability to identify chemicals with the potential to interact with the endocrine system.

Federal agencies are also instructed by the ICCVAM Authorization Act of 2000 to ensure that new and revised test methods are valid prior to their use.

D. Validation Process

While this paper focuses on the criteria for assay validation, it is useful to review the validation process because a common understanding of these concepts is helpful to understanding the discussion in this paper.

In general, EPA is following the five-part or stage validation process outlined by the Interagency Coordinating Committee for the Validation of Alternative Methods (ICCVAM) (NIEHS, 1997). EPA believes that it is essential to recognize that this process was specifically developed for *in vitro* assays that were intended to replace *in vivo* assays. The fundamental problem confronting the U.S. EPA is how to adapt and work with this process for a far wider range of rodent and ecological *in vivo* assays ranging from simple, lower tier screens to higher tier multigenerational reproductive and developmental tests for Tier II.

The first stage of the process outlined by ICCVAM is *test development*, an applied research function which culminates in an initial protocol. As part of this phase, EPA prepares a Detailed Review Paper (DRP) to explain the purpose of the assay, the context in which it will be used, and the scientific basis upon which the assay's protocol, endpoints, and relevance rest. The DRP reviews the scientific literature for candidate protocols and evaluates them with respect to a number of considerations, such as whether the candidate protocols meet the assay's intended purpose, the costs and other practical considerations. The DRP also identifies the developmental status and questions related to each protocol; the information needed answer the questions; and, when possible, recommends an initial protocol for the initiation of *prevalidation* in which the protocol is refined, optimized, and initially assessed for transferability and performance. Several different types of studies are conducted during the assay's prevalidation phase depending upon the state of development of the method and the nature of the questions that the protocol raises. The initial assessment of transferability is generally a trial in a second laboratory to determine that another laboratory besides the lead laboratory can follow the protocol and execute the study. *Inter-laboratory validation* studies are conducted in independent laboratories with the protocol that was optimized during prevalidation. The results of these studies are used to determine inter-laboratory variability and to set or cross-check performance criteria. Inter-laboratory validation

is followed by *peer review*, an independent scientific review by qualified experts, and by *regulatory acceptance*, adoption for regulatory use by an agency. ICCVAM also recognizes that the validation process may not be able to supply complete information on the performance of the assay.

Strict adherence to this process is not necessary for a study to be determined scientifically validated. The European Centre for the Validation of Alternative Methods (ECVAM) has proposed a modular approach to validation. The modular approach regards validation in the context of the data needed to demonstrate relevance and reliability, i.e., satisfy the validation criteria, rather than as a linear process (i.e., prevalidation followed by validation). This also means that where data exists for a test method prior to undertaking prevalidation or validation programs, these studies or modules should not be needlessly repeated. This is often referred to as using existing or retrospective data (i.e., data generated prior to and outside of the initiation of the validation program). There are seven data modules: *test definition, within laboratory variability, protocol transferability, between laboratory variability, predictive capacity, applicability domain, and minimum performance standards*. Data modules can be filled in any sequence with existing (i.e., retrospective) data or data obtained prospectively from the validation program (Hartung, 2004).

The Organization for Economic Cooperation and Development (OECD) employs a phased approach to the inter-laboratory validation of assays in their Test Guidelines Program (TGP) that does not follow the strict division between prevalidation and validation. If a standardized protocol exists, Phase I is an inter-laboratory study with strong positive chemicals to demonstrate that laboratories can successfully execute the standardized protocol. If no standardized protocol exists, Phase I begins with a sub-phase in which the protocol is standardized. Phase II is an inter-laboratory study generally conducted with weaker substances and one or more negative substances to determine the performance characteristics of the assay and generally provides some information on the reproducibility of the assay over time. Several of the EDSP assays are included in the OECD TGP; EPA intends to rely on the OECD process for validation of these EDSP assays.

E. A Historical Perspective

Test guidelines, especially in the ecotoxicity area, have been tested before their regulatory use in a paradigm that focused on reliability. Ecotoxicologists focused on standardizing the protocol and demonstrating and evaluating the ability of laboratories to correctly execute the protocol and measure its endpoints. This interlaboratory study, typically known as a ring test, was run with several laboratories using one or more positive compounds. Relevance was assumed: the test species were themselves members of the ecosystem of interest and, therefore, were considered to be of direct relevance to ecological hazard and risk assessment. While the test species served as surrogates for other species in the environment, the question of extrapolation to other species was considered as a matter to be addressed in risk assessment, not test method validation.

F. Validation Criteria

Criteria for the validation of alternative test methods *(in vitro* methods designed to replace animal tests in whole or in part) have generally been agreed upon in the U.S. by ICCVAM, in Europe by the ECVAM, and internationally by the OECD. These criteria as stated in by ICCVAM (NIEHS, 1997) are as follows:

1. The scientific and regulatory rationale for the test method, including a clear statement of its proposed use, should be available.
2. The relationship of the endpoints determined by the test method to the *in vivo* biologic effect and toxicity of interest must be addressed.
3. A formal detailed protocol must be provided and must be available in the public domain. It should be sufficiently detailed to enable the user to adhere to it and should include data analysis and decision criteria.
4. Within-test, intra-laboratory and inter-laboratory variability and how these parameters vary with time should have been evaluated.
5. The test method's performance must have been demonstrated using a series of reference chemicals, preferably coded to exclude bias.
6. Sufficient data should be provided to permit a comparison of the performance of a proposed substitute test to that of the test it is designed to replace.
7. The limitations of the test method must be described (e.g., metabolic capability).
8. The data should be obtained in accordance with Good Laboratory Practices (GLPs).
9. All data supporting the assessment of the validity of the test methods including the full data set collected during the validation studies must be publicly available and, preferably, published in an independent peer reviewed publication.

OECD Guidance Document 34 and the Solna Principles list only eight criteria as they combine criteria 7 and 2. As noted, these validation criteria were developed for alternative methods, and the OECD Guidance Document 34 on validation of test methods (OECD 2005) notes the need for flexibility and adaptability in applying validation criteria in many contexts and circumstances that will be encountered in the validation of different assays. Guidance Document 34 states:

> A set of principles for validation, also called the "Solna Principles", were developed at an OECD Workshop in Solna Sweden in 1996, where it was agreed that these Principles apply to the validation of new or updated test methods for hazard assessment, whether they are *in vivo* or *in vitro*, or tests for effects on human health or the environment. The extent to which these validation principles are addressed will vary with the purpose, nature, and proposed use of the test method. There are differences between *in vivo* assays and *in vitro* or *ex vivo* assays which should be considered in applying the validation principles. (Guidance Document 34, paragraphs 12-13)
>
> The principles for validation apply to all test methods. Scientific rigour is always required, regardless of the scope of the validation, the type of test method, or whether the method is new, revised or historical. The amount and kind of information needed and the criteria applied to a new test method depends on a number of factors. These include:
>
> - the regulatory and scientific rationale for the use of the test method,
> - the type of test method being evaluated (e.g., new test, existing test)
> - the proposed use of the test method (mechanistic adjunct, screening, definitive, replacement test, etc.)

- the proposed applicability domain of the test method (restricted chemical classes, organic chemicals that are not polymers, etc.)
- the relationship of the test species to the species of concern,
- the mechanistic basis of the test and its relationship to the effect of concern, and
- the history of use of the test method, if any, within the scientific and regulatory communities. (Guidance Document 34, paragraphs 30-31)

ICCVAM also states that the extent to which validation criteria are met will vary with the method and its proposed use and that the validation of tests for different types of effects requires different approaches (NIEHS, 1997).

II. Application of the Validation Criteria to the EDSP

This section addresses how the EPA's EDSP generally interprets and plans to apply the validation criteria discussed in section I.F. above to the validation of the assays proposed for the EDSP. EPA regards validation as an assessment of the utility and limitations of an assay to serve a given purpose and peer review as an audit of the underlying scientific evidence being assessed. Regulatory acceptance of an assay as part of the EDSP will be decided by the EPA after peer review.

The proposed use of a method in a regulatory scheme sets the standard for what must be demonstrated during validation because a test method is validated for a specific purpose. The purpose of assays in Tier I is to function as a comprehensive screen to identify chemicals with the potential to interact with the estrogen, androgen or thyroid hormone systems and these assays will be validated for this purpose. Validating an assay to be a sensitive and reliable screen is different and less burdensome than the validation of an assay to predict effects in an intact organism. In addition, the Tier I assays are intended to function as part of a battery so that the limitations of one assay are offset by the strengths of another. Validation should clarify the strengths and weaknesses of each assay so that the proper mix of assays can be selected for the EDSP Tier I screening battery. Tier II assays identify adverse effects resulting from exposure to endocrine disruptors and provide a quantitative estimate of the amount of a test chemical necessary to cause an adverse response.

The application of some criteria does not vary as a function of an assay. Criteria 1-3 and 7-9, as outlined in section I.F., are generally applicable to all kinds of validation without modification. Criterion 4 will need some adaptation for Tier II tests. These criteria require an explanation of the nature of the test, its proposed use, and the endpoints being measured; the availability of the protocol; an evaluation of variability; the availability of all data supporting validation; compliance with GLP; and peer review.

Criterion 5 (demonstration of test method performance with coded reference chemicals) represents the biggest challenge for many of the assays in the EDSP. EPA believes that application of this criterion is highly dependent on the type of assay and the endpoints being validated.

For replacement tests, Criterion 6 is based on the need to assess the correlation of the results

between the older and the replacement test to demonstrate their strict and broad equivalence. Since none of the assays in the EDSP are replacement tests, Criterion 6 (comparison of the performance of a proposed substitute test to that of the test it is designed to replace) does not apply at this time. While it is of interest to know how well these screens perform in identifying chemicals that are positive in Tier II tests, this can only be done to a limited extent at this time. Examples of this type of assessment have been conducted with the uterotrophic and Hershberger *in vivo* screens against other *in vivo* data including multi-generational tests. However, the real proof of the performance of the Tier I screens will be a retrospective comparison of the performance of the battery with Tier II results after sufficient Tier II data have been generated in the testing program. This is why EPA is committed to a retrospective analysis of the test data generated on the first 50 to 100 chemicals tested in the EDSP.

Table 1 summarizes the applicability of the validation criteria to the EDSP.

Validation Criteria	**Applicability**
1. The scientific and regulatory rationale for the test method, including a clear statement of its proposed use, should be available.	Applies to all assays without modification.
2. The relationship of the endpoints determined by the test method to the *in vivo* biologic effect and toxicity of interest must be addressed.	Applies to all assays without modification.
3. A formal detailed protocol must be provided and must be available in the public domain. It should be sufficiently detailed to enable the user to adhere to it and should include data analysis and decision criteria	Applies to all assays without modification.
4. Within-test, intra-laboratory and inter-laboratory variability and how these parameters vary with time should have been evaluated.	Applies to all *in vitro* screening level assays without modification. For *in vivo* assays, biological variability is a large component of the variability, but the ability of several laboratories to measure novel endpoints will be evaluated. Variability over time will not generally be evaluated for in vivo assays.
5. The test method's performance must have been demonstrated using a series of reference chemicals preferably coded to exclude bias.	The nature of the assay will influence how this criterion is applied. For example, if the relevance of the assay is known and the question to be answered is sensitivity, a relatively small set of well chosen chemicals can be used to obtain this information.
6. Sufficient data should be provided to permit a comparison of the performance of a proposed substitute test to that of the test it is designed to replace.	This criterion applies to replacement tests only.
7. The limitations of the test method must be described (e.g., metabolic capability).	This applies to all assays without modification.
8. The data should be obtained in accordance with Good Laboratory Practices (GLPs).	Applies to all assays; however, it is recognized that it may be necessary to employ some non-GLP labs for validation in special circumstances.
9. All data supporting the assessment of the validity of the test methods including the full data set collected during the validation studies must be publicly available and, preferably, published in an independent peer reviewed publication.	Applies to all assays without modification

Improvements and replacements for the first generation assays will be considered at a later date and some, such as the recombinant ER and AR binding assays, are under development now. The approach taken to validate future assays as replacements will likely depend upon how closely they resemble the original assay. For some closely related methods, a demonstration that assays meet performance criteria established for the assay they are replacing may be sufficient.

A. Demonstrating Relevance

Flexibility and adaptability are recognized as essential in the application of the Guidance Document 34 criteria, but they are especially important for criteria relating to the demonstration of relevance. Relevance can be based on three factors:

- scientifically accepted theory (Criteria 1 and 2),
- empirical demonstration of test performance (test data to establish a correlation between endpoints and effects)(Criteria 5 and 6), and
- direct observation of inherently relevant endpoints (Criterion 2).

The third factor is not applicable to alternative tests, but it is covered by Criterion 2 which requires that the relationship of the endpoints determined by the test method to the *in vivo* biologic effect and toxicity of interest be described. The contribution of each of these three factors to establishing the relevance of an assay differs according the assay being validated.

1. The Role of Scientific Understanding

The scientific rationale for a test method is the scientific understanding upon which the method is based. For endocrine disruptors, the scientific rationale for a test rests upon an understanding of the endocrine system and how external substances can interact with it. When scientific rationale for a test method is based on well-accepted scientific theory, it can provide robust support for an assay's relevance. This can reduce the burden of empirical proof to establish relevance per se, as there is no need to provide additional justification for well-accepted scientific principles. For example, there is no need to prove that receptor binding is the mechanism by which the endocrine system functions, and that mimicry of the hormone or interference with its binding to the receptor is a potential way to interfere with the function of the endocrine system. Similarly, the more closely the test method's endpoint is to the biological effect of interest, the less need there is to demonstrate relevance by empirical means (OECD, 2005). But the opposite is true as well: when an assay is based on novel principles or a limited understanding of the basis on which it works, or its relevance to the biological system or endpoint of interest is not well substantiated, a more complete and robust empirical demonstration of relevance is required.

The scientific rationale for the test method and an understanding of the relationship of a test method's endpoints to the biologic effect will serve as the primary support for the relevance of the many EDSP assays to endocrine disruption. There is substantial understanding of the endocrine system and how it functions, and unlike replacement assays, which can be compared

to existing assays to gauge scientific meaningfulness and usefulness, endocrine assays have relatively few reference materials and, thus, must rely more heavily on scientific understanding of the endocrine system. For this reason, EPA believes that the description of that rationale and description of the test method's endpoints to the biologic effect should typically be held to higher standards than for "alternative" assays which augment their rationales with a more complete empirical demonstration of relevance. For alternative tests, which by definition do not directly measure the toxicity of interest, the relationship of a test method's endpoint and the biological effect or toxicity of interest is expressed in the form of a prediction model. For *in vivo* assays where direct observations of toxicity are made, there is no need for a prediction model, but guidance on data interpretation is provided.

For Tier I screens, the effects of interest are based on the known ways in which chemicals can affect the endocrine system: effects on hormone synthesis, receptor binding, interaction with the hypothalamic-pituitary axis, interference with hormone transport, alterations in hormone metabolism, and organism responses regulated by hormones. Assays that detect or measure these effects are relevant to a determination that a chemical does or does not have potential to affect the endocrine system, and data that demonstrate that an assay performs one of these functions will usually be sufficient empirical support for the relevance of the assay.

2. Empirical Demonstration of Relevance

For assays that are replacing other assays (i.e., alternative test methods), determining assay performance by testing reference chemicals is conceptually simple, but critical to validation. Through its use, the existing test has an established data base that can be used as a reference data set for validation of the replacement assay. Known positive and negative agents are often easily identifiable as data exist from the older assay to be replaced. Therefore, the new replacement assay is tested with a representative subset of the chemicals tested in the original assay (ranging from positive to negative) and the results or predictions of toxicity obtained with the new assay are compared with the results obtained in the old assay. For *in vitro* methods that replace animal tests, the results of a prediction model, which converts the *in vitro* result into a prediction of *in vivo* toxicity, are compared with the results found in the original test. If the predictions made by the replacement test are good enough for its intended purpose—it is as good or better than the original test if it is a total replacement—the assay can be said to be validated for its intended purpose.

In contrast, for new screens or tests such as those being developed by the EDSP there is no comparable reference data base. For some assays such as the receptor binding, uterotrophic, and Hershberger assays that have an established history of use with pharmacological compounds there will be data on compounds that impact the endocrine system that may be tested during the development and validation phases of an assay. The key need in these cases is to demonstrate that these assays are capable of detecting weak acting, commercial chemicals that may mimic the pharmacological compounds when administered at high doses; however, the data related to the endocrine activity of commercial chemicals and pesticides compounds is quite limited requiring a careful review of the literature and selection of compounds for the validation program. For example, besides the strong natural and synthetic estrogens, sufficient evidence of estrogenic

activity existed for nonylphenol, methoxychlor, genistein, and *o,p'*-DDT to use these chemicals in the validation of the uterotrophic assay.

For many assays in the EDSP, relevance is based to a substantial extent on biological or mechanistic understanding of the assay (e.g., receptor binding) and/or direct observation of the endpoint of interest (e.g., fish multi-generation reproductive toxicity test). This is supplemented by a demonstration that the assay is measuring the endpoint of interest and, in the case of screens, has the sensitivity or ability of the assay to detect weakly active chemicals using reference chemicals with known endocrine activity. The selection of the reference chemicals is critical as they become the design target for the screens and test whether an assay will meet its regulatory purpose. Ideally, the reference chemicals should provide some diversity in potency, chemical structure, and properties; however, it must be recognized that the ability to include representative chemicals is constrained both by the limited number of chemicals that can be tested (financial and practical considerations) and the chemicals for which endocrine activity is known. Screens in the EDSP will typically be judged to be adequately sensitive if they identify the benchmark chemicals in the reference chemical library in multi-laboratory studies. For tests, such as modification of the multi-generational protocol, a few chemicals in a single laboratory are adequate because directly observable adverse effects are being measured and significant time and resources are involved.

The guidance that the set of reference chemicals should be representative of the full domain of applicability of the test addresses the limitations of the assay (Criterion 7) more than proof of relevance for non-alternative assays since the expected result of most of these substances will not be known. While it may be feasible to test a broader set of chemicals for this purpose, it may not be cost effective to always do so, and it is infeasible to do so for *in vivo* tests. In addition, in many cases it may not be necessary to test a broad range of chemicals to determine the limitations of an assay. For instance, we know that cytotoxicants cannot be tested in cell-based assays, so an assay for cell viability is normally included as part of such a test system to safeguard against false positives due to cytotoxicity. Similarly, substances that denature proteins will interfere with receptor binding assays and assays of enzyme activity such as the aromatase assay. ECVAM has also considered resources in selecting a reference set of approximately a dozen chemicals consisting of several strong positives, weak positives, and negatives for the validation of their reproductive effects screens, so there are precedents for what EPA is proposing.

Reference chemicals in the EDSP will generally be based on research in which the EPA or others have tested chemicals and developed an understanding as to their mode of action of on the basis of test results confirmed in two or more well-run independent studies. Typically, a larger number of chemicals will be run in a single laboratory than in interlaboratory validation studies to conserve both animals and funds. Since a limited number of chemicals are well studied for endocrine effects, the reference set will usually be composed of a small number of chemicals, and as a consequence, it is likely that many of the same chemicals will be used during prevalidation and in the interlaboratory validation studies.

3. Direct Observation of Inherently Relevant Endpoints

Tier II assays identify adverse effects resulting from interference with the endocrine system and provide a quantitative estimate of the amount of a test chemical necessary to cause an adverse response. All Tier II assays encompass the reproductive cycle and early maturation stages of organisms of various selected taxa because these life stages are known to be most sensitive to regulation by the endocrine system. The biological effects/toxicities of interest for the Tier II tests encompass measures of reproductive competence, growth, and development that are known to be controlled by the endocrine system. The relevance of these assays is based upon the direct observation of inherently relevant endpoints; thus, no empirical confirmation of relevance is needed for these assays.

B. Reliability

An assay will generally be considered to be reliable by EPA if its overall variability is low enough to give a level of sensitivity or power[1] consistent with the purpose the assay is intended to serve. The power of the assay depends on the variability of the assay (baseline noise), the magnitude of the positive response (strength and magnitude of signal), and the number of replicate test units per treatment level. For screening level assays, this purpose is to provide information to determine whether higher tier "definitive" studies should be conducted or not, and success can be judged as to whether the assay detects effects on the selected benchmark or reference chemicals. In screening assays, test concentrations or dose levels can be adjusted to maximum exposure or tolerated levels to increase the sensitivity of an assay in the face of higher variability in an assay endpoint. In definitive tests, test concentrations or doses are expected to be at the margins of effect where endpoint variability is more influential on test sensitivity.

EPA is evaluating three types of variability—within-test, between tests in the same laboratory (intra-laboratory), and between laboratory (inter-laboratory) variability—in its program to validate assays for the EDSP. Some preliminary data on these parameters will be obtained during prevalidation; however, the primary purpose of the inter-laboratory validation studies is to generate this information to judge the performance of the standardized protocol as it pertains to the observed results in comparison to the expected results for each endpoint. Most *in vitro* studies will be run in triplicate at the same time and in the same laboratory. Thus, within-test variability for these studies is measured by the variability across three replicates. *In vivo* studies specify a certain number of replicate test units (i.e., individuals, litters, breeding pairs, tanks, pens, etc.) per treatment level. Thus, within-test variability of *in vivo* studies reflects the variability of responses observed among the replicate test units within a given treatment level. Some of the observed variability will be due to the inherent biological variability of the test system. Sources of run-to-run variability include a number of protocol and laboratory factors such as reagent preparation, pipetting, or other factors that may not be constant over time. Variability is also strongly influenced by laboratory proficiency, including experience in conducting the assay. Assessing the performance of the assay in one or more inexperienced labs may give the Agency a preview of how well the assay's range of performance upon its initial

[1] Power is the probability that a statistical significance level will reject the null hypothesis (no treatment effect) for the alternative hypothesis (that treatment caused an effect). Thus, power is a measure of the ability of a test to detect an effect, given that the effect actually exists. See Appendix A for a more complete explanation of power.

regulatory implementation, but if the data are to be used to set reasonable performance criteria—benchmarks of performance that should be realized by proficient laboratories—laboratories should have some training and opportunity to become proficient before variability data are collected and used to set performance criteria. Three to five laboratories will be typically considered sufficient to generate these data in interlaboratory studies.

Some *in vivo* assays have only one endpoint; others have several endpoints for the same mode of action; still others are multi-modal and contain one or more endpoints for each mode of action. Variability will be determined for each endpoint measured. Endpoints that show such high variability as to be relatively insensitive to detect the effects of the test chemical may be dropped from the assay or made optional in the final protocol. The validation study plan and the final validation report will discuss what measures are being made and how they are being compared (i.e., what statistical analyses are being performed). Reference chemicals for interlaboratory studies will be selected to test the reliability of an assay as discussed in the next section.

C. Tailoring Validation Studies to Different Types of EDSP Assays

It is useful to organize the following discussion around the following assay types: *in vitro* assays, single mode of action *in vivo* assays (e.g., the uterotrophic assay, which detects estrogenic effects *in vivo*), and multi-modal *in vivo* assays (e.g., the pubertal female assay, which detects effects on the HPG axis, estrogen, thyroid, and steroidogenesis.). Assays within each of these categories share certain characteristics which influence the degree of flexibility in the application of the validation criteria that is both necessary and appropriate.

In vitro single mode-of-action screening assays are not intended to be replacement assays. EDSP *in vitro* assays are expected to detect or measure the interaction of chemicals with certain components of the endocrine system, e.g. receptor binding or enzyme inhibition. They complement *in vivo* assays in a battery, not but do not replace them. They will be validated for this intended purpose, not as replacement assays. Therefore, a strict correlation or comparison with *in vivo* results would **not** be the determining factor in judging their validation. Although it is expected that there will be a relatively high correlation between *in vitro* and *in vivo* data, there are a number of reasons (e.g., absorption, distribution, metabolism, and excretion) why *in vivo* and *in vitro* results might diverge.

Receptor binding is a good example of EPA's approach for validating a single mode-of-action *in vitro* assay. EPA's confidence in the relevance of receptor binding assays rests heavily on the understanding—including mathematical models—that has been developed over the past 50 years on competitive binding. Knowing that the theory of receptor binding is well established, the logic of the validation of receptor binding assays can be outlined as follows:

1) Demonstrate that the assay is truly measuring competitive binding. This is established by running a saturation curve and measuring Bmax and Kd.

2) Test a small set of chemicals that specifically bind to the receptor with differing potency. Ten chemicals should provide the range of potency that is required to

demonstrate the ability of the assay to detect weak binders (i.e. those with IC 50's between 1 uM and 1 mM). Several negative chemicals will demonstrate the assay's ability to discriminate between positive and negative chemicals.

3) Derive performance standards for the assay to ensure repeatability by competent laboratories and a means to identify laboratories that are not performing the assay adequately. The performance standards will typically be values that are well established for the controls.

4) Develop procedures for the interpretation of data from the model. These data interpretation procedures will define what constitutes a negative result, an equivocal result, and a positive result in the binding assay.

5) Measure the reliability of the assay (variability within a run, within a lab between runs, and between labs) during an inter-laboratory validation study. The variability of the assay will have a direct bearing on the ability of the assay to distinguish positive responses.

Similar considerations apply to competitive inhibition of the enzyme aromatase.

Although sufficient to demonstrate that the assays are functioning as intended, 10 chemicals are insufficient to perform a statistically meaningful analysis of all of the indicators of accuracy of the assay. EPA has shown this through a Monte Carlo simulation that at least 10-25 chemicals are necessary for the prediction of some of these parameters (sensitivity, specificity, positive predictivity, and negative predictivity) and 100 or more are needed for others.[2] Reference chemicals are limited to those for which there are reliable data and availability. The performance of these assays will be judged on the basis of these assays to discriminate between moderate, weak, and negative chemicals.[3] Ten chemicals are also insufficient to establish the limitations of

[2] The Monte Carlo simulation addressed the precision of the estimates of sensitivity, specificity, positive predictivity, negative predictivity, and concordance (sometimes referred to as Cooper statistics) as a function of the number and choice of reference chemicals sampled from the domain of applicability of an assay. This analysis illustrated the change in precision of the estimates as the sample size increased and provides and indication of the numbers of reference chemicals needed to estimate the Cooper statistics with high precision. It shows that a large number of reference chemicals (100-200), divided among true positive and true negative chemicals, are necessary in order to have meaningful estimates of all of the performance parameters; however, depending on circumstances, some parameters may be estimated with as few as 10-25 chemicals. The precision of the sensitivity estimate depends on the true sensitivity of the assay and number of true positive chemicals in the sample. The precision of the specificity estimate similarly depends on the true specificity of the assay and the number of true negative chemicals. Thus, for assays with high specificity and sensitivity, the number of chemicals needed for precise estimates is smaller than for assays with lower sensitivity and specificity, but in all the cases considered, 50 or more true positive and 50 or more true negative chemicals should be included in the reference chemical data base. The underlying true positive and negative predictivity is additionally a function of the prevalence of positive chemicals in the domain of applicability. All other things being equal, positive predictivity would be expected to be lower and negative predictivity would be expected to be higher if the prevalence of true positive chemicals in the population is lower (Battelle, 2005).

[3] As an example chemical strengths we can regard strong binders as chemicals that bind with a log IC50 within 2 orders of magnitude of the natural ligand (e.g., RBA > 1), moderate binders as those with an IC50 approximately 3-4 orders of magnitude higher than the IC50 of the natural ligand, and weak binders as those with log IC50 within 2-3

these assays; however, as noted in section II.A. 2, testing a large number of chemicals is not always necessary to determine the limitations of the assay.

Single mode-of-action *in vivo* screening assays are next in terms of complexity. Examples include the uterotrophic assay (for estrogenicity), the Hershberger (for androgenicity and anti-androgenicity), and the frog metamorphosis assay (for effects on thyroid). Like the receptor binding and aromatase assays discussed above, these assays are meant to play a defined role in a Tier I battery for assessing a single mode of action such as estrogenicity or effects on a single target organ such as the thyroid. (It is recognized that there are several modes of action by which the thyroid can be acted on, but an OECD Detailed Review Paper on thyroid test procedures concluded that no mechanistically based thyroid tests are ready for inclusion in a validation program.) Single mode-of-action *in vivo* screening assays are not meant to replace any other assay, and are intended to complement the other assays in the battery where information may be missing or equivocal or where redundancy is warranted. All three of these assays are being validated through the OECD TGP.

It is not feasible to test as many chemicals using *in vivo* assays as with *in vitro* assays because of animal welfare, expense, time, and the limited number of reference chemicals on which there are reliable data. Ideally, a limited number of chemicals (e.g., 6 to 10) exhibiting a range of responses expected (strong to negative) will be used to validate single-mode-of-action *in vivo* assays. These chemicals will demonstrate the ability of the laboratories to obtain reproducible data with chemicals of varying potency and indicate the ability of the assay to discriminate between positives, negatives, and chemicals of varying physiochemical and metabolic properties. This will demonstrate whether the screening assay meets the basic criterion: the ability to detect chemicals that interact with the endocrine system. In practice, the use of a limited number of chemicals will place great emphasis on the consideration of chemical candidates and the transparent communication of the rationale for their selection.

Multiple-mode-of-action *in vivo* screening and definitive assays are the most complex assays to validate. All other Tier I assays, such as the pubertal assays and fish reproductive screen, are intended to be multimodal or cover several modes of action. Thus, they are more apical in nature as are the Tier II assays.

These multimodal screens and Tier II tests are generally conducted according to the standard *in vivo* toxicological paradigm: a negative control group (sham or vehicle-treated control) is used, and multiple dose levels of test chemical are administered. It is not practical to provide a positive control for each individual mode of action since this would result in a huge increase in the use of animals and costs with relatively little information gained in return.

Multimodal *in vivo* assays are included in Tier I because only whole animal assays can serve as a model that incorporates all aspects of the endocrine system: control of hormone production and feedback control through the hypothalamic–pituitary axis, enzymes for the synthesis of

---

orders of magnitude of the limit dose. Negative chemicals are those for which an IC50 cannot be determined within the limit dose.

hormones, secretion, transportation mechanisms through the blood, and receptors and response elements in target tissues.

For Tier I multimodal screens, each basic mode of action fundamental to the estrogen, androgen or thyroid pathways will be tested with known positive substances during the course of validation of a screening assay, usually in a single lab during the prevalidation phase. A chemical that is negative by all modes of action will also be tested when possible either in prevalidation or during the interlaboratory validation study. However, chemicals proven to be negative by all modes of action may be difficult to identify because relatively few chemicals have been tested in the battery of relevant screens and tests. In addition, negative results are frequently not reported in the literature. When a general negative chemical cannot be found, in some cases, it may be satisfactory to find one that is negative in one sex (e.g., an antiandrogen in a female) or positive in one mode of action but negative in others (e.g., a thyroid active chemical that is negative with respect to the estrogen and androgen systems). To compensate for this limitation, the Agency proposes to reassess the performance of Tier I multimodal assays several years after implementation to compare the performance of the Tier I battery with Tier II outcomes. ICCVAM has recognized that judgments of validation status may change over time as new scientific information about a test method is acquired (NIEHS, 1997).

For Tier II definitive tests, validation will not focus on testing each mode of action but will include an appropriate chemical to evaluate each endpoint so that data on endpoint variability can be obtained across laboratories. For Tier II tests, it may be appropriate to include certain targeted studies to validate specific endpoints to assess variability instead of the full-scale Tier II tests. Such shorter-term and smaller-scale evaluations could address specific endpoint variability issues more easily and practically than the full-scale tests, but a single full-scale study may be necessary to demonstrate that the protocol is practical and that all endpoints can be effectively measured in a single study. These shorter term tests could be one-generation tests or even screening assays if they have common endpoints.

Coding of samples for analysis is employed during validation studies to remove investigator bias, where such bias may occur. Certain direct organismal observations, like organ weight or number of eggs produced, are not subjective and, therefore, are not appreciably influenced by this bias. However, certain indirect measures and other observations, like behavior or morphology scoring could are more subjective and steps should be employed to eliminate bias. There are, however, practical limitations to conducting a fully blind test. In aquatic tests, for example, the investigator must have knowledge of the chemical and the various treatment groups to effectively monitor the exposure concentrations and maintain the test system. In another case, histopathological analysis is inherently subjective, but full coding is also impractical. The most common suggestion from non-pathologists is to analyze the slides blind, but this is universally rejected by practicing pathologists. Pathologists have well established procedures for reading slides that will allow them to recognize unfamiliar or novel pathologies and still read the slides without bias (Crissman et al., 2006). Although the nature of the test compound may be blinded to the pathologist, the slides are read beginning with control specimens followed by the experimental high-dose group. Typically, the remaining lower-dose groups are examined only if pathology is seen in the high-dose group. After all slides have been read, apparent treatment-

related pathologies are confirmed by a blind re-reading of the slides or by a second pathologist. Therefore, the degree to which coding or blinding is necessary for a validation trial is dependent on the test type and the endpoints employed.

Effects seen in whole animal studies in well-conducted independent replicates are relevant and reliable as markers or effects for the test species. To what other species within their taxonomic group (i.e., fish, birds, mammals) they are relevant is a separate question, and the uncertainty of this extrapolation is addressed during hazard and risk assessment. For human health effects, human data are generally cited as the gold standard, but sufficient quantities of high quality human data almost never exist and cannot be ethically obtained for most endpoints of interest in toxicological testing, so this suggestion is mainly theoretical, not practical. For ecotoxicity testing, while it is possible to develop data in some target species, it is clearly not feasible to do so for very many species—for reasons of resources, availability of species, and ability to raise certain species under laboratory conditions. Thus, the Tier II assays in the EDSP are being validated as model systems: species applicability will be presumed and extrapolation across species will be addressed in the risk assessment process, not as part of validation.

In this area, we must for now content ourselves with the philosophy of Aristotle:

> It is the mark of an instructed mind to rest satisfied with the degree of precision which the nature of the subject permits and not to seek an exactness where only an approximation of the truth is possible.
> Aristotle

III. Peer Review

It is EPA's policy that major scientific and technically based work products related to Agency decisions be peer-reviewed. According to EPA's Science Policy Council Handbook on Peer Review (U.S. EPA, 2000),

> Peer review is a process for <u>enhancing</u> a scientific or technical work product so that the decision or position taken by the Agency, based on that product, has a sound, credible basis…. Effective use of peer review is indispensable for fulfilling the EPA mission and therefore deserves high-priority attention from program managers and scientists....

For completeness the following table lists the assays being considered for the EDSP. It is expected that not all assays listed below will undergo peer review. Some assays will undergo peer review as part of an OECD validation effort, others after a successful validation effort as part of the Agency's program will undergo peer review, and still others may not survive the validation process so that a peer review will not be necessary. For assays undergoing peer review, EPA will prepare a Summary Validation Report which will summarize all of the data relevant to the validation of the assay and demonstrate how the validation was achieved.

Tier I Assays	Tier I Assay Battery	Tier II Assays
Pubertals (M & F) Adult Male Fish Screen Frog Metamorphosis* AR Binding (RPC) rrAR Binding * ER Binding (RUC) hrER Binding* Aromatase Steroidogenesis Hershberger** Uterotrophic**	Battery To be Determined	Two-generation Mammalian † Two-generation Avian* Two-generation Fish* Two-generation Mysid* Amphibian Growth and Reproduction*

* It is not clear at this time whether EPA or OECD will be responsible for this peer review.
** OECD peer review.
† This assay would not be subject to peer review but is included for completeness in the listing of assays in the EDSP

A. Tier I Assays

The mechanism that will be used to peer review Tier I assays will be an EPA peer review contract. For each assay, the contractor will compile a list of qualified peer review candidates who are independent of those who performed the work or who have been involved in the development or refinement of the protocol, including those who have provided EPA with expert advice throughout the validation process. The potential peer reviewers will be identified from among academia, government, and private sector institutions, based on their subject matter expertise, availability, and lack of conflict of interest or past involvement in the project. From this pool of candidate reviewers, the contractor will establish a "balanced" peer review panel consisting of approximately 5 peer reviewers. The contractor will provide the reviewers with the integrated validation summary report and any supporting documentation, such as study reports, that are needed for the peer review, along with a list of charge questions that will be developed by EPA.

The contractor will compile the peer review record which will include the peer review document and all supporting materials given to the peer reviewers; the instructions/charge to the peer reviewers; all comments, information, and materials received from the peer reviewers; public comments; meeting summary; and names, affiliations, qualifications of the peer review panel members. EPA will use the peer review record to make a final determination as to a Tier I assay's suitability for inclusion in the Tier I battery, and finalize the assay for implementation, if determined to be acceptable. EPA plans to begin peer reviewing Tier I assays by late-2007. This schedule is dependent upon the successful completion of studies that are currently underway.

B. Tier I Assay Battery

Subsequent to peer review of individual assays and prior to initiating testing, EPA intends to

propose a battery of Tier I screening assays to be peer reviewed by EPA's Science Advisory Panel (SAP), with participation of EPA's Science Advisory Board (SAB). While the exact format for the SAP/SAB review has not yet been determined, it is expected that the proposed battery along with the materials supporting its composition will be provided to a panel of approximately 15 to 20 reviewers. Some of the panel members may be individuals who participated in review of one or more Tier I assays, and some individuals will be new to the EDSP peer review process. Use of some of the same reviewers for both the Tier I assays and the Tier I battery is intended to ensure that individuals familiar with the individual assays are represented when the battery is discussed. This should not present a conflict of interest because the context of the review and the questions being asked of the battery reviewers will differ from what is asked of the Tier I assay reviewers (e.g., questions posed to the SAP/SAB reviewers would pertain to whether the proposed battery adequately covered the endpoints of interest for estrogen, androgen, and thyroid while questions posed to the Tier I assay reviewers would focus on whether or not the particular assay was sufficiently validated).

C Tier II Assays

The peer review strategy for the Tier II assays will be development after EPA has experience with peer review of the Tier I battery. New assays will have a full SAP/SAB review. Modified versions of current assays may have a more limited form of peer review depending upon the scope of the modifications and enhancements. At present, no modifications have been made to the two-generation mammalian assay although OECD and EPA are developing a one-generation alternative to the traditional two-generation assay.

IV. Summary and Conclusions

This paper has outlined the approach EPA is using in validating assays for the EDSP. Some validation criteria are more important for assays in the EDSP than they are for alternative assays and vice versa. Some validation criteria apply to all assays universally; others will need to be adapted to fit the assay being validated. The following statements summarize the conclusions reached in Section II.

- Relevance for an assay and its endpoint(s) can be based on three factors—scientifically accepted theory, empirical demonstration of test performance (data generated during a validation program that correlates the performance of the test with an authoritative reference value), and direct observation of inherently relevant endpoints. The contribution of each factor differs according the assay and type of endpoints being validated.
- The case for the relevance of many assays in the EDSP is based on well-accepted scientific theory and an understanding of the relationship of the test method's endpoints to the biologic effect.
    - When the scientific rationale for a test method is based on well-accepted scientific theory, it can provide robust support for the assay's relevance, and the need for empirical proof to establish relevance is lessened. Most endocrine modes of action

and the endpoints that are responsive targets are relatively well known; thus, the Agency has confidence in these circumstances that the validation of these assays can be accomplished by using a carefully selected, limited set of chemicals.
    - The more closely the test method's endpoint is to the biological effect of interest, the less need there is to demonstrate relevance by empirical means.
    - The description of the scientific rationale and relationship of the test method's endpoints to the biologic effect should generally be held to higher standards when they are the primary support for the relevance of an assay.
- Data generated during the validation program will address the reliability of the assay and provide added evidence of its relevance. The variability of an assay will generally be considered satisfactory by EPA if it is low enough to give a level of sensitivity or power consistent with the purpose the assay is intended to serve.
    - The primary role of empirical data in addressing relevance is to demonstrate the sensitivity of the assay and, to some degree, the specificity.
    - The role of negative chemicals in the validation of assays in the EDSP is to demonstrate that the assay can discriminate between positive and negative chemicals.
    - Given the amount of resources for *in vivo* assays versus *in vitro* assays, it is impractical to use large numbers of chemicals. This often will preclude the calculation of statistically meaningful estimates of sensitivity and specificity.
- The Tier II assays in the EDSP are being validated as model systems: species applicability will be presumed and extrapolation across species will be addressed in the risk assessment process, not as part of validation.
- Comparison of the new test with the test it is designed to replace does not apply at this time to assays in the EDSP since they are all new tests.
- Science is dynamic. Experience gained through regulatory use of the assays, such as screening the first group of 50-100 chemicals, will generate far more data than can be generated through any validation program. It may enhance confidence in the assays or prompt a reanalysis of its validation status. New assays that are more efficient and effective will replace older assays as science progresses and additional data become available.

Glossary:

Assay: any laboratory test procedure ; includes both Tier I screens and Tier II tests.

Positive predictivity: the probability that an outcome is truly positive when the test result is positive. The positive predictivity is a function of the sensitivity of the test and fraction of true negative chemicals in the population.

Negative predictivity: the probability that an outcome is truly negative when the test result is negative. The negative predictivity is a function of the selectivity of the test and fraction of true positive chemicals in the population.

Relevance: whether a test is meaningful and useful for a particular purpose.

Reliability: the reproducibility of results from an assay within and between laboratories and over time.

Screen: a relatively short in vitro or in vivo assay designed to identify a chemical for further evaluation.

Sensitivity: the ability of an assay to detect positive chemicals. It is defined mathematically as the ratio of positive outcomes in the test divided by the number of true positives.

Specificity: the ability of an assay to detect negative chemicals. It is defined mathematically as the ration of negative outcomes in the test divided by the number of true negatives.

Tier I: a battery of screening level assays designed to detect chemicals that may affect the estrogen, androgen or thyroid hormone systems.

Tier II: multigenerational tests in different taxa designed to identify and quantify the adverse effects of chemicals that interact with the estrogen, androgen and thyroid systems. However, since these tests measure a range of reproductive and developmental parameters that may be affected by other factors than the endocrine system, they are not by themselves necessarily diagnostic of endocrine disruption.

Validation: the process by which the reliability and relevance of a test method are evaluated for a particular use.

References:

Crissman JW; Goodman DG; Hildebrandt PK; Maronpot RR; Prater DA; Riley JH; Seaman WJ; Thake DC. 2004. Best Practices Guidelines: Toxicologic Pathology. *Toxicologic Pathology*. Pp. 126-131.

Hartung T et al. A Modular Approach to the ECVAM Criteria on Test Validity. ATLA **32**, 467-472, 2004.

National Institute of Environmental Health Sciences. "Validation and Regulatory Acceptance of Toxicological Test Methods, A Report of the ad hoc Interagency Coordinating Committee on the Validation of Alternative Methods." Research Triangle Park, NC. NIH Report 97-3981. March, 1997.

Organisation for Economic Cooperation and Development. Final Report of the OECD Workshop on Harmonization of Validation and Acceptance Criteria for Alternative Toxicological Test Methods. August , 1996.

Organisation for Economic Cooperation and Development. Guidance Document on the

Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment. Guidance Document No. 34. June 2005.

U.S. EPA. Science Policy Council Handbook: Peer Review, 2nd Edition. Office of Science Policy, U.S. Environmental Protection Agency, Washington, DC. EPA 100-B-00-001.