

November 9, 2005

## Validation of Assays in the EDSP

### I. Introduction

#### A. Issue

This paper describes how commonly accepted validation criteria are interpreted and applied to the assays being validated by the U.S. EPA for use in the Endocrine Disruptor Screening Program (EDSP).

#### B. 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 accepted 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. 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 the 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).

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 validation process outlined by the Interagency Coordinating Committee for the Validation of Alternative Methods (ICCVAM) (NIEHS, 1997). The first 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 criteria upon which the assay rests. 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 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 to be 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). 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 data or data obtained prospectively (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. 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. Validation Criteria**

Criteria for the validation of alternative test methods (*in vitro* methods designed to replace animal tests) have generally been agreed upon in the U.S. by ICCVAM, in Europe by the ECVAM, and internationally by the OECD. These criteria are as follows (OECD, 1996; NIEHS, 1997):

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.

As noted, these validation criteria were developed for alternative methods, and the OECD Guidance Document 34 on validation of test methods (OECD 2005) stresses the need for flexibility in applying validation criteria stating:

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.

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 EPA's EDSP generally interprets and plans to apply the validation criteria discussed in Unit I.E. above to the validation of the assays of 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 is the decision of the Agency to include an assay as part of the EDSP.

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 system and these assays will be validated for this purpose. Validating an assay to be a sensitive and reliable screen is different and substantially 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 assays. Criteria 1-4 and 7-9 are generally applicable to all kinds of validation without modification. 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 being validated.

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. 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 they meet performance criteria established for the assay they are

replacing may be sufficient.

### **A. Demonstrating Relevance**

Flexibility is recognized as essential in the application of the criteria, but it is 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 (Criteria 5 and 6), and
- direct observation of inherently relevant endpoints.

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 to 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 thereby reducing the burden of empirical proof to establish relevance as there is no need to reprove 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 interferes 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 the assay is based on novel principles or a limited understanding of the basis on which it works or of its relevance to the biological system or endpoint of interest, 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 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.

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 all measures of reproductive competence and physical and behavioral 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.

## **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 easily identifiable. 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 where a substantial reference data base does not exist and for practical reasons cannot be generated, there is no gold standard set of reference chemicals with which to compare the assay. For the EDSP, relevance is based mainly on biological or mechanistic understanding of the assay and/or direct observation of the endpoint of interest, and reference chemicals with known endocrine activity provide information demonstrating that the assay is measuring the endpoint of interest and the sensitivity or ability of the assay to detect weakly active chemicals. The selection of these chemicals is critical as they become the design target for the assay and test whether the 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 limited both by the limited number of chemicals that can be tested and the chemicals for which endocrine activity is known (many will be pharmaceutical products developed for a specific endocrine mode of action). Screens in the EDSP will typically be judged to be adequately sensitive if they identify correctly the benchmark chemicals in the reference chemical library in single-laboratory studies or in multi-laboratory studies.

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 on the basis of test results confirmed in two or more well-run independent studies. A larger number of chemicals will typically be run in a single laboratory than in interlaboratory validation studies to conserve both animals and funds. Since few 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. 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 most like the *in vitro* replacement assays because *in vitro* assays are often used as alternative or replacement tests. However, unlike replacement assays, EDSP *in vitro* assays are expected to directly measure relevant endpoints that complement *in vivo* assays in a battery, not replace them. They will be validated for this intended purpose, not as replacement assays, and therefore, 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 (such as absorption, distribution, metabolism, and excretion) why *in vivo* and *in vitro* results might diverge.

Receptor binding assays will generally be tested with known receptor binders of various types and potencies (including some negative chemicals) to demonstrate that the assay is effective in detecting specific binding to the receptor and to determine the assay's ability to detect weak binders (i.e. those with IC<sub>50</sub>'s between 1 uM and 1 mM). 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. Similar considerations apply to competitive inhibition of the enzyme aromatase. Approximately 10 reference chemicals are being tested in the interlaboratory studies of the ER and AR binding assays and the aromatase inhibition assays. 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: 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.<sup>1</sup> 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.

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. They are not meant to replace any other assay, but to complement the other assays in the battery. 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. A few chemicals (*e.g.*, 6 to 10) exhibiting a range of responses expected (strong, medium, weak, and negative) will be tested in 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 different strength. This will demonstrate whether the screening assay meets the basic criterion: the ability to detect chemicals that interact with the endocrine system—in this case by a particular mode of action. While qualitative answers to this question would be adequate, most of these assays will give quantitative information which could be also used to set priorities for Tier II testing should that be necessary.

Multiple-mode-of-action *in vivo* screening and definitive assays are the most complex assays to validate. All Tier II assays and some Tier I assays, such as the pubertal assays and fish reproductive screen, are multimodal. These tests are generally conducted according to the standard *in vivo* toxicological paradigm: only a negative control (sham or vehicle-treated control)

---

<sup>1</sup> 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 an 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).

and multiple dose levels of test chemical are administered. It is not practical to provide a positive control for each of the modes of action since that would result in a huge use of animals and increase in the cost of tests 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 integrates all aspects of the endocrine system: control of hormone production through the hypothalamic–pituitary axis, enzymes for the synthesis of hormones, secretion, transportation mechanisms through the blood, and receptors and response elements in target tissues. For Tier I screening assays, 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 find because so few chemicals have been tested using relevant tests and 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 performance of Tier I multimodal assays will be reassessed 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 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.

A special note needs to be added about coded chemicals for testing in aquatic organisms and birds with treated diet. This is generally not done for aquatic studies as it necessary to monitor the amount of chemical to which aquatic organisms are actually exposed. Actual chemical concentrations can deviate from target concentrations by diluter error, volatilization, hydrolysis or adsorption to surfaces. Analytical chemical procedures must be tailored to the chemical, so the analytical chemist and diluter technicians, at a minimum, must know the identity and target concentration of the test material. In addition, diluters in some laboratories are designed such that concentration assignments cannot really be blind. This complicates any blind testing procedure, but in some laboratories separation of function (chemists and technicians versus biologists) may permit the conduct of effectively blind studies even if the identity of the chemical is known to some personnel involved.

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 they are relevant is a

separate question. 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

## **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 consistent with the purpose the assay is intended to serve. The power of the assay depends on the variability of the assay, the magnitude of the positive response, and the number of replicate test units per treatment level. For screening level assays, this purpose is to provide ‘suggestive’ 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 doses can be adjusted to maximum exposure or tolerated levels that 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 test in the same laboratory, and between 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. 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 which is due to biological variability. Variability from run to run may reflect a number of other factors such as reagent preparation, pipetting, or other factors that may not be constant over time. Variability is also strongly influenced by laboratory competence including experience in conducting the assay. Using untrained labs may give the Agency a preview of how well the assay will be conducted upon its initial 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. Three to five laboratories will be typically considered sufficient for 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 analysis is being performed). Reference chemicals for interlaboratory studies will be selected to test the reliability of an assay as discussed in the previous section.

### 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 enhancing 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, and still others may not survive the validation process. At present no modifications have been made to the two-generation mammalian assay. For assays undergoing peer review, EPA will prepare a Summary Validation Report (Appendix A to be added) which will summarize all of the data relevant to the validation of the assay and demonstrate how the validation was achieved.

<u>Tier I Assays</u>	<u>Tier I Assay Battery</u>	<u>Tier II Assays</u>
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.
- \*\* A peer review has been conducted by OECD but its outcome is being questioned.
- † 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**

It is anticipated that 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 final validation report and any supporting documentation that is needed for the peer review, along with a list of charge questions that will be developed by EPA.

The panel will review and comment on the assay and meet in a public forum in which the public will have an opportunity to comment. 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 mid-2006. 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 is currently under development. 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.

## **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 because they are the primary evidence of the relevance of the assays in the EDSP; others play a less significant role for assays in the EDSP than they do for alternative assays and may be applicable only with adaptation. The following statements summarize the conclusions reached in Section II.

- Relevance can be based on three factors—scientifically accepted theory, empirical demonstration of test performance, and direct observation of inherently relevant endpoints; the contribution of each factor differs according the assay being validated.
- The case for the relevance of assays in the EDSP is based primarily on well-accepted scientific theory and an understanding of the relationship of the test method’s endpoints to the biologic effect.
  - When 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.
  - 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.
  - The primary role of empirical data in addressing relevance is to demonstrate the sensitivity of the assay.
  - 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.
  - The practical numbers of chemicals to be used in most EDSP screening assays preclude the calculation of statistically meaningful estimates of sensitivity, specificity.
- 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 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 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.

## **References:**

Battelle. Sensitivity of Cooper Statistics to the Number and Choice of Chemicals Used in Validation Study [unpublished analysis prepared for EPA] March 21, 2005.

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, 2<sup>nd</sup> Edition. Office of Science Policy, U.S. Environmental Protection Agency, Washington, DC. EPA 100-B-00-001.