Regulatory Toxicology and Pharmacology 72 (2015) 552–561

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

Analysis of EPA's endocrine screening battery and recommendations for further review

Adam W. Schapaugh [a,*], Lisa G. McFadden [b], Leah M. Zorrilla [c], David R. Geter [c], Leah D. Stuchal [d], Neha Sunger [e], Christopher J. Borgert [d,f]

[a] Monsanto Company, St. Louis, MO, USA
[b] The Dow Chemical Company, Midland, MI, USA
[c] Bayer CropScience, Research Triangle Park, NC, USA
[d] Center for Environmental and Human Toxicology, College of Veterinary Medicine, University of Florida, Gainesville, FL, USA
[e] Department of Health, West Chester University, PA, USA
[f] Applied Pharmacology and Toxicology, Inc., Gainesville, FL, USA

ARTICLE INFO

Article history:
Received 14 April 2015
Available online 2 June 2015

Keywords:
Assay
Endocrine Disruptor Screening Program
Tier 1 Endocrine Screening Battery
Hormone system

ABSTRACT

EPA's Endocrine Disruptor Screening Program Tier 1 battery consists of eleven assays intended to identify the potential of a chemical to interact with the estrogen, androgen, thyroid, or steroidogenesis systems. We have collected control data from a subset of test order recipients from the first round of screening. The analysis undertaken herein demonstrates that the EPA should review all testing methods prior to issuing further test orders. Given the frequency with which certain performance criteria were violated, a primary focus of that review should consider adjustments to these standards to better reflect biological variability. A second focus should be to provide detailed, assay-specific direction on when results should be discarded; no clear guidance exists on the degree to which assays need to be re-run for failing to meet performance criteria. A third focus should be to identify permissible differences in study design and execution that have a large influence on endpoint variance. Experimental guidelines could then be re-defined such that endpoint variances are reduced and performance criteria are violated less frequently. It must be emphasized that because we were restricted to a subset (approximately half) of the control data, our analyses serve only as examples to underscore the importance of a detailed, rigorous, and comprehensive evaluation of the performance of the battery.

1. Introduction

Section 408(p) of the 1996 Food Quality Protection Act (FQPA) (US EPA, 1996) mandated the US Environmental Protection Agency (EPA) to develop and maintain a screening program to investigate the potential of chemicals to interfere with the endocrine system in humans. After passage of the FQPA, EPA convened a federal advisory committee, the Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC), to assess the current state-of-the-science and assist in developing an endocrine screening program (The Keystone Center, 1996). EDSTAC concluded that the assays needed to evaluate endocrine activity varied significantly in their degree of development and recommended that EPA: (1) adopt a two-tiered, hierarchical testing and evaluation framework; and (2) initiate a research program to develop, standardize, and validate the necessary assays (EDSTAC, 1998).

Based largely on the EDSTAC recommendation, EPA developed a two-tiered framework for the Endocrine Disruptor Screening Program (EDSP). The EDSP Tier 1 Endocrine Screening Battery (EDSP ESB) consists of eleven assays (five *in vitro* and six *in vivo*) intended to identify the potential of a chemical to interact with the estrogen, androgen, thyroid, or steroidogenesis systems in humans and wildlife. EDSP Tier 2 testing is designed to identify and establish a quantitative, dose–response relationship for any adverse effects that might result from interactions with the endocrine system. Thus, the purpose of the EDSP ESB is to differentiate chemicals with the potential to interact from those that have little or no such potential. EPA intends to use a weight-of-evidence approach to make this determination (Borgert et al. 2011a; US EPA, 2011a).

EPA began issuing the first set of test orders for the EDSP in 2009. Following the recommendation of a joint committee of the

[*] Corresponding author at: Monsanto Company, 800 North Lindbergh Blvd., St. Louis, MO 63167, USA. Fax: +1 314 694 8525.
E-mail address: adam.schapaugh@monsanto.com (A.W. Schapaugh).

http://dx.doi.org/10.1016/j.yrtph.2015.05.028
0273-2300/© 2015 The Authors. Published by Elsevier Inc.

EPA Science Advisory Board and Scientific Advisory Panel (SAB/SAP) to initially screen 50–100 chemicals (US EPA, 1999), EPA has elected to proceed in phases, with the first phase composed of 67 chemicals (hereafter, List 1) (US EPA, 2009a). Recognizing that many of the eleven EDSP ESB assays had not been fully validated, the SAB/SAP also recommended that EPA review all testing methods once data were collected from the initial 50 to 100 chemicals (US EPA, 1999). Indeed, any evaluation leading to Tier 2 testing would be premature until the battery has been thoroughly evaluated (Borgert et al., 2011b).

EPA has required manufacturers and importers of List 1 chemicals to conduct the EDSP ESB in accordance with specific guidance (Table 1); out of the first 67 test orders, data for 52 chemicals have been submitted to EPA. We have collected data for control groups from approximately half of those submitted to the agency. Our purpose in doing so was to gather information in support of a critical review of the EDSP ESB. One focus of that review was to quantify the normal biological ranges for key endpoints that had insufficient historical databases when the test guidelines were written. Without knowledge of the magnitude of endpoint variances, a reliable weight-of-evidence determination will be difficult, especially if only 1 or 2 significant differences are observed across the battery. A second focus of the review was on the appropriateness of performance criteria defined in the test guidelines and the potential to revise and improve them. Finally, we examined the impact of methodological differences between laboratories on assay and endpoint variances. It must be emphasized that because the review was restricted to a subset of the List 1 control data, our analyses serve only as examples to underscore the importance of a detailed, rigorous, and comprehensive evaluation of the performance of the battery.

2. Methods

Efforts to validate EDSP ESB assays have proven to be difficult and time-consuming (Borgert et al., 2011b). Prior to the completion of List 1 studies, data were insufficient to conduct robust performance evaluations of many of the screens. Our approach sought to leverage data generated by companies complying with List 1 test orders in an experiential manner to more thoroughly evaluate critical parameters that reflect assay performance and biological variance. It is beyond the scope of the paper to summarize in detail each of these assays, but a general overview is provided below. Complete information is provided by EPA at: http://www.epa.gov/ocspp/pubs/frs/publications/Test_Guidelines/series890.htm.

Table 1
EPA's EDSP Tier 1 Endocrine Screening Battery.

	Test guideline[†]	Control datasets
In vitro screens		
Estrogen Receptor (ER) Binding Assay	890.1250	13
ER Transcriptional Activation Assay	890.1300	15
Androgen Receptor Binding Assay	890.1150	16
Aromatase Assay	890.1200	23
Steroidogenesis Assay	890.1550	17
In vivo screens		
Uterotrophic Assay	890.1600	26
Hershberger Assay	890.1400	21
Male Pubertal Assay	890.1500	25
Female Pubertal Assay	890.1450	25
Fish Short-Term Reproductive Assay	890.1350	22
Amphibian Metamorphosis Assay	890.1100	33

[†] Test Guidelines can be accessed at: http://www.epa.gov/ocspp/pubs/frs/publications/Test_Guidelines/series890.htm.

2.1. Assay summary and purpose

2.1.1. In vitro screens

2.1.1.1. Estrogen Receptor (ER) Binding Assay. The ER Binding Assay responds to chemicals that interact with the estrogen receptor. The assay measures the receptor-binding affinity of a chemical by evaluating its ability to displace the endogenous hormone 17β-estradiol, which is prepared from rat uterine cytosol. Substances that bind the estrogen receptor in rats are presumed to be capable of binding the estrogen receptor in humans because the binding domain of the receptor is highly conserved across species. Note that the assay does not discern whether a chemical will potentially act as an agonist or antagonist.

2.1.1.2. Estrogen Receptor Transcriptional Activation (ERTA) Assay. The ERTA Assay responds to chemicals that bind and activate the estrogen receptor. In the assay, estrogen enters the target cell, binds the receptor, and activates a signaling pathway leading to the production of a luciferase enzyme. The luciferase product is measured by the addition of luciferin; this causes a light-emitting reaction to occur. The amount of light emitted is proportional to the potency of the estrogen. Consequently, this procedure can measure the ability of a chemical to induce hERα-mediated transactivation of luciferase gene expression. This assay is specific for potential agonist activity.

2.1.1.3. Androgen Receptor (AR) Binding Assay. The AR Binding Assay responds to chemicals that interact with the androgen receptor. The assay measures the receptor-binding affinity of a chemical by evaluating its ability to displace a bound reference androgen isolated from rat ventral prostates. Substances that bind the androgen receptor in rats are presumed to be capable of binding the androgen receptor in humans because there is a high degree of DNA sequence conservation in the receptor across mammalian phylogenetic lines. The assay does not discern whether a chemical will potentially act as an agonist or antagonist.

2.1.1.4. Aromatase Assay. The Aromatase Assay responds to chemicals that inhibit the catalytic activity of the cytochrome P450 enzyme aromatase. The assay is conducted by incubating human recombinant microsomes with multiple concentrations of the test chemical, radio-labeled androstenedione, and a cofactor for a specified period of time. The rate of tritiated water released during the conversion of the androstenedione to estrone is measured and used as an indicator of aromatase activity.

2.1.1.5. Steroidogenesis Assay. The Steroidogenesis Assay allows direct measurement of hormone production, cell viability, and cytotoxicity in the human H295R adrenocortical carcinoma cell line. The assay is performed under standard culture conditions in 24-well plates. After an acclimation period of 24 h, the cells are exposed to increasing concentrations of the test chemical in triplicate for 48 h. After the period of exposure, the levels of testosterone and estradiol released by the cells are measured.

2.1.2. In vivo screens

2.1.2.1. Male Pubertal Assay. The Male Pubertal Assay responds to chemicals with potential anti-thyroid, androgenic, and anti-androgenic activity, as well as agents that alter pubertal development through mechanisms that induce changes in gonadotropins, prolactin, or through alterations in hypothalamic function. It can also respond non-specifically in ways that confound interpretation of endocrine mechanisms (Marty et al., 2011). The assay uses juvenile male rats that are exposed to the test chemical daily by oral gavage from post-natal day (PND) 23 through 53. This duration of treatment is needed to detect potential changes in pubertal development and effects on the thyroid gland. The assay evaluates age at preputial separation (PPS), a marker of pubertal development in the male, changes in weights of accessory sex glands [testes, epipidymides, ventral prostate, levator ani/bulbocavernosus muscle (LABC), seminal vesicle with coagulating gland, (SVCG)], changes in weights of the pituitary, liver, adrenals, kidneys, thyroid, and hormone concentrations [thyroxine (T4), TSH (thyroid stimulating hormone), and testosterone].

2.1.2.2. Female Pubertal Assay. The Female Pubertal Assay responds to chemicals with potential anti-thyroid, estrogenic, and anti-estrogenic activity, as well as agents that alter pubertal development through mechanisms that induce changes in luteinizing hormone, follicle stimulating hormone, prolactin, or through alterations in hypothalamic function. As seen with the Male Pubertal Assay, it can also respond non-specifically in ways that confound interpretation of endocrine mechanisms (Marty et al., 2011). The assay uses juvenile female rats that are exposed to the test chemical daily by oral gavage from PND 22 through 42. This duration of treatment is needed to detect potential changes in pubertal development and effects on the thyroid gland. The assay evaluates age at Vaginal Opening (VO), a marker of pubertal development in the female, changes in weights of uterus, ovaries, pituitary, liver, adrenal, kidneys, and thyroid, as well as changes in hormone concentrations (T4, TSH), and estrous cyclicity.

2.1.2.3. Amphibian Metamorphosis Assay (AMA). The Amphibian Metamorphosis Assay assesses growth and developmental endpoints as well as thyroid histopathology of African clawed frog (*Xenopus laevis*) tadpoles in response to test chemicals delivered via the water. The rate of tadpole metamorphosis is controlled, in part, by the Hypothalamus–Pituitary–Thyroid (HPT) axis. Thus, the AMA assesses chemicals for the potential to interfere with the normal function of the HPT axis. The assay represents a generalized model to the extent that it is based on the conserved structures and functions of the HPT axis in vertebrates. The AMA is the only assay in the EDSP ESB that detects potential effects on the thyroid gland in an animal undergoing morphological development.

2.1.2.4. Fish Short-Term Reproductive Assay (FSTRA). The Fish Short-Term Reproductive Assay responds to chemicals that may interfere with the normal function of the Hypothalamus–Pituitary–Gonadal (HPG) axis. The assay uses sexually mature fathead minnow (*Pimephales promelas*) to detect changes in spawning, morphology, gonadal histopathology, and specific biochemical endpoints that reflect disturbances in the HPG axis. The FSTRA represents a generalized model to the extent that it as based on the conserved structures and functions of the HPG axis in vertebrates.

2.1.2.5. Uterotrophic Assay. The Uterotrophic Assay responds to chemicals with estrogenic activity. The assay utilizes either an animal model in which the HPT axis is not functional (immature) or an ovariectomized model, leading to low baseline uterine weights and low levels of endogenous estrogen. Uterine weights (with and without uterine fluids) of animals exposed to the test chemical are compared to a vehicle control group; a statistically-significant increase in uterine weights (wet or blotted) indicates a positive response. In addition, a positive control group, administered ethinyl estradiol, is included in the assay.

2.1.2.6. Hershberger Assay. The Hershberger Assay identifies chemicals with androgenic or anti-androgenic activity. The assay uses castrated male rats and measures the weights of five androgen-dependent tissues (Cowper's Gland, LABC, ventral prostate, SVGC, and glans penis) to determine the potential for androgen agonists, antagonists, and 5α-reductase inhibitors. A positive result in the agonist assay is a statistically-significant increase in 2 or more of the androgen-dependent tissues compared to the vehicle control group. A positive result in the antagonist assay is a statistically-significant decrease in 2 or more of the androgen-dependent tissues compared to the control group (administered vehicle control plus testosterone propionate (TP, 0.2 or 0.4 mg/kg). Positive controls are utilized in both the agonist (vehicle control + TP) and antagonist assay (TP + flutamide).

2.2. Analysis

Access to a subset of the control data was obtained from companies complying with List 1 test orders. The number of datasets, by assay, is shown in Table 1. Prior to analysis, all test chemical reference names were removed and several parameters were coded, including the submitter, laboratory that performed the assay, company name, microsome source (if unique), non-standard reference chemicals, and non-standard solvents. In addition, datasets were made uniform by standardizing units.

The first focus of the analysis was to quantify the normal biological ranges for required endpoints in each assay. To do so, summary statistics (overall mean, standard deviation, coefficient of variation [CV], and minimum and maximum raw values) were calculated for each endpoint across studies. The second focus of the analysis was on the appropriateness of performance criteria and the potential to improve them with the evaluation of data from a large number of studies. The EDSP ESB test guidelines (see Table 1) include performance criteria for control data, which laboratories should meet to confirm that the assays were conducted properly. Fitted curve parameters, mean values and CVs must fall within the defined ranges for the assay results to be considered fully acceptable. The number and percentage of studies failing these criteria were tabulated for each assay. The final focus of the analysis was on the impact of methodological differences between laboratories on endpoint variances and whether performance violations were systemic (observable across laboratories) or driven by a minority of participating laboratories. Mean ranges and average and maximum variances and CVs (where appropriate) were calculated for each endpoint. The following linear model was fit to evaluate the impact of laboratory on endpoint variance:

$$y_{ij} = \mu + l_i + \varepsilon_{ij} \tag{1}$$

where y_{ij} is the observed coefficient of variation for the jth study conducted by the ith laboratory; μ is the overall mean; l_i is the fixed effect of the ith laboratory; ε_{ij} is the residual error.

The MIXED PROCEDURE in SAS (SAS, 2012) was used to fit model (1) separately for each endpoint with at least 25% of studies in violation of the maximum CV or had CVs averaged across studies that exceeded the maximum CVs allowed by the test guidelines. Note the analysis was weighted to account for the different numbers of studies submitted by each laboratory.

3. Results

For clarity and ease of presentation, this section has been organized by assay. Endpoint names are presented in *italicized* font. Space for tables in the main text has been reserved for reporting the number and percentage of studies failing the performance criteria. All summary statistics for the analysis of the List 1 control data can be found in the Supplementary material.

3.1. In vitro screens

3.1.1. Estrogen Receptor (ER) Binding Assay

The ER Binding Assay uses 17β-estradiol and norethynodrel as strong and weak positive controls, respectively. Non-linear regression methods are used to fit curves (separately for the strong positive control, weak positive control, and test chemical) to a modified version of the Hill Equation (Hill, 1910). Fitted curve parameters, which include percent binding at the top and bottom of the curve, an estimate of the Hill Slope, and the natural logarithm of the residual standard deviation, must fall within defined tolerance limits for the assay results to be considered fully acceptable. The number and percentage of studies failing these criteria are presented in Table 2. Summary statistics are provided in Tables A1 and A2. Of the thirteen studies that reported estimates of 17β-estradiol fitted curve parameters, two (15%) had a least one tolerance limit violation. In addition, six studies (46%) reported at least one parameter estimate for norethynodrel outside of its acceptable range. Three studies (23%) were in violation of the tolerance limit for *Bottom (% Binding)*, while four studies (31%) failed the criterion for *Hill Slope* $(\log_{10}(M)^{-1})$.

3.1.2. Estrogen Receptor Transcriptional Activation (ERTA) Assay

The ERTA Assay uses 17β-estradiol, 17α-estradiol, and 17α-methyltestosterone as a strong agonist, weak agonist, and very weak agonist, respectively. Non-linear regression methods are used to fit curves (separately for the strong agonist, weak agonist, very weak agonist, and test chemical) to Hill's Logistic Equation (Hill, 1910). Fitted curve parameters include estimates of the Hill Slope and concentration that provokes a response midway between the baseline and maximum responses (EC$_{50}$). These parameters, along with calculations of 10 and 50 percent of control (PC) values, must fall within defined tolerance limits for the assay results to be considered fully acceptable. The number and percentage of studies failing these criteria are presented in Table 3. Summary statistics are provided in Tables A3 and A4. A majority (62% and 67%) of fitted curve parameters failed to fall within the tolerance limits for the very weak agonist. In addition, the ranges of acceptable values for *Hill Slope* $(\log_{10}(M)^{-1})$ were problematic for the strong agonist and weak agonist.

3.1.3. Androgen Receptor (AR) Binding Assay

The AR Binding Assay uses methyltrienolone (R1881) and dexamethasone as strong and weak positive controls, respectively. Non-linear regression methods are used to fit curves (for the strong positive control, weak positive control, and test chemical, separately) to a modified version of the Hill Equation (Hill, 1910). Fitted curve parameters, which include percent binding at the top and bottom of the curve, and an estimate of the Hill Slope, must fall within defined tolerance limits for the assay results to be considered fully acceptable. The number and percentage of studies failing these criteria are presented in Table 4. Summary statistics are provided in Tables A5 and A6. All fitted curve parameters were within the tolerance limits for the weak positive control. A single run was in violation of the tolerance limit for *Hill Slope* $(\log_{10}(M)^{-1})$ for R1881. No other violations were observed.

Table 2
Performance criteria in the Estrogen Receptor Binding Assay.

Criterion	Tolerance limit(s)	Number of runs	Number (%) failed
Radioinert 17β-estradiol fitted curve parameters			
log$_e$(Residual Std. Dev.)	⩽2.35	18	0 (0%)
Top (% Binding)	94–111	19	1 (5%)
Bottom (% Binding)	−4 to 1	19	1 (5%)
Hill Slope (log$_{10}$(M)$^{-1}$)	−1.1 to −0.7	19	1 (5%)
Weak positive control (norethynodrel) fitted curve parameters			
log$_e$(Residual Std. Dev.)	⩽2.35	18	0 (0%)
Top (% Binding)	94–111	19	0 (0%)
Bottom (% Binding)	−4 to 1	19	3 (16%)
Hill Slope (log$_{10}$(M)$^{-1}$)	−1.1 to −0.7	19	4 (21%)

Note: Performance criteria based on control substances only. Refer to OCSPP Test Guideline 890.1250 for a full listing of performance requirements.
n = 13 studies, 1–3 runs reported per study.

Table 3
Performance criteria in the ERTA.

Criterion	Tolerance limit(s)	Number of runs	Number (%) failed
17β-Estradiol			
log(PC$_{50}$)	−11.4 to −10.1	37	1 (3%)
log(PC$_{10}$)	<−11	37	0 (0%)
log(EC$_{50}$)	−11.3 to −10.1	37	1 (3%)
Hill Slope	0.7–1.5	37	6 (16%)
17α-Estradiol			
log(PC$_{50}$)	−9.6 to −8.1	37	0 (0%)
log(PC$_{10}$)	−10.7 to −9.3	37	0 (0%)
log(EC$_{50}$)	−9.6 to −8.4	37	1 (3%)
Hill Slope	0.9–2.0	37	10 (27%)
17α-Methyltestosterone			
log(PC$_{50}$)	−6.0 to −5.1	21	13 (62%)
log(PC$_{10}$)	−8.0 to −6.2	33	22 (67%)

Note: Performance criteria based on control substances only. Refer to OCSPP Test Guideline 890.1300 for a full listing of performance requirements. *Abbreviations:* PC$_{10}$ = 10 percent of control; PC$_{50}$ = 50 percent of control; EC$_{50}$ = half maximal response.

3.1.4. Aromatase Assay

The Aromatase Assay uses 4-hydroxyandrostenedione (4-OH ASDN) as a positive control. Non-linear regression methods are used to fit curves (for the positive control and test chemical, separately) to a modified version of the Hill Equation (Hill, 1910). Fitted curve parameters, which include percent binding at the top and bottom of the curve, an estimate of the slope, and an estimate of the concentration corresponding to 50% of control activity (IC$_{50}$), must fall within defined tolerance limits for the assay results to be considered fully acceptable. In addition, performance guidelines have been defined for the mean aromatase activity in the absence of an inhibitor and the mean background control activity. The number and percentage of studies failing these criteria are presented in Table 5. Summary statistics are provided in Tables A7 and A8. A majority (17 out of 23; 74%) of studies had at least 1 run in violation of the tolerance limits for the fitted curve

Table 4
Performance criteria in the Androgen Receptor Binding Assay.

Criterion	Tolerance limit(s)	Number of runs	Number (%) failed
R1881 fitted curve parameters			
Top (% Binding)	82–114	34	0 (0%)
Bottom (% Binding)	−2.0 to 2.0	34	0 (0%)
Hill Slope (log$_{10}$(M)$^{-1}$)	−1.2 to −0.8	34	1 (3%)
Weak positive control (dexamethasone) fitted curve parameters			
Top (% Binding)	87–106	34	0 (0%)
Bottom (% Binding)	−12 to 12	34	0 (0%)
Hill Slope (log$_{10}$(M)$^{-1}$)	−1.4 to −0.6	34	0 (0%)

Note: Performance criteria based on control substances only. Refer to OCSPP Test Guideline 890.1150 for a full listing of performance requirements.

Table 5
Performance criteria in the Aromatase Assay.

Criterion	Tolerance limit(s)	Number of runs	Number (%) failed
4-OH ASDN fitted curve parameter			
Slope	−1.2 to −0.8	67	3 (4%)
Top (%)	90–110	67	10 (15%)
Bottom (%)	−5 to 6	67	3 (4%)
log(IC_{50})	−7.3 to −7.0	67	17 (25%)
Minimum aromatase activity (nmol/mg-protein/min)	⩾0.1 nmol/mg protein/min	61	2 (3%)
Background control activity (nmol/mg-protein/min)	⩽15% of full activity control	48	2 (4%)

Note: Performance criteria based on control substances only. Refer to OCSPP Test Guideline 890.1200 for a full listing of performance requirements. *Abbreviations:* 4-OH ASDN = 4-hydroxy androstenedione; IC_{50} = concentration corresponding to 50% of control.

parameters. In addition, 7 studies (30%) had at least 3 runs in violation of the tolerance limits and 1 study had at least 5 violations. Two parameters were responsible for most of the violations: 15% and 25% of all runs failed the criteria for *Top (%)* and *log(IC_{50})*, respectively.

3.1.5. Steroidogenesis Assay

The Steroidogenesis Assay utilizes a known inducer (Forskolin) and known inhibitor (Prochloraz) to determine if expected changes in hormone production are detectable. Levels of testosterone and estradiol must fall within defined ranges for the assay results to be considered fully acceptable. The number and percentage of studies failing these criteria are presented in Table 6. Summary statistics are provided in Tables A9 and A10. Nearly half (41%) of all runs violated the acceptable level of *Estradiol* production for the positive control inhibitor. One-quarter (25%) of runs reported basal production of *Estradiol* below the acceptable minimum of 40 pg/mL. For all other measures, hormone production was within acceptable levels at least 96% of the time.

3.2. In vivo screens

3.2.1. Male Pubertal Assay

The Male Pubertal Assay utilizes a vehicle control as a comparator to the test chemical. Performance criteria have been established for a number of endpoints; these criteria are intended to indicate whether the assay was sufficiently sensitive to allow conclusions on whether the test chemical did, or did not, affect pubertal development and growth. The number and percentage of studies failing the performance criteria defined in the test guidelines are presented in Table 7. Developmental and growth endpoints are summarized in Tables A11 and A12. All (25 out of 25) studies had at least 1 mean-range violation and at least 1 CV violation. In addition, 9 (36%) studies had at least 3 mean-range violations and 21 (84%) had at least 3 CV violations. Finally, 1 (4%) study had at least 5 mean-range violations and 8 (32%) had at least 5 CV violations.

Several endpoints, including *Body Weight (g) at Preputial Separation, Final Body Weight (g), Seminal Vesicle + Coagulating Gland (SVCG) w/Fluid (g), Epididymis, left (g), and Epididymis, right (g)*, had CVs averaged across studies that exceeded the maximum CVs allowed by the test guidelines (Tables 7 and A12). The overall mean of *Kidneys (g)*, reported as 2.09 g in Table A11, was outside of the acceptable mean range of 2.42–3.05 g (Table 7), and 76% of studies fell outside of the acceptable range. One study was outside the acceptable mean range for *Age (PND) at Preputial Separation* and 2 studies were outside the acceptable range for *Body Weight (g) at Preputial Separation*. More importantly, 9 (36%) studies exceeded the maximum CV for *Age (PND) at Preputial Separation* and 22 (88%) studies exceeded the maximum CV for *Body Weight (g) at Preputial Separation*. In addition, 9 (36%) studies failed the mean range for *Serum TSH* and 7 studies (28%) failed the CV criterion.

Model (1) was fitted to endpoints with at least 25% of studies in violation of the maximum CV or had CVs averaged across studies that exceeded the maximum CVs allowed by the test guidelines. Eight endpoints (Table 8) met one or both of these criteria. Based on the overall *F*-tests, there were no statistically-significant

Table 6
Performance criteria in the Steroidogenesis Assay.

Criterion	Tolerance limit(s)	Number of runs	Number (%) failed
Minimum basal production			
Testosterone	⩾500 pg/mL	16	0 (0%)
Estradiol	⩾40 pg/mL	16	4 (25%)
Induction (10 μM Forskolin)			
Testosterone	⩾1.5-fold solvent control	47	1 (2%)
Estradiol	⩾7.5-fold solvent control	47	0 (0%)
Inhibition (1 μM Prochloraz)			
Testosterone	⩽0.5-fold solvent control	46	0 (0%)
Estradiol	⩽0.5-fold solvent control	37	15 (41%)

Note: Performance criteria based on control substances only. Refer to OCSPP Test Guideline 890.1550 for a full listing of performance requirements.

Table 7
Performance criteria in the Male Pubertal Assay.

Endpoint	Acceptable mean range	Number (%) of studies failed	Maximum CV (%)	Number (%) of studies failed
Age (PND) at preputial separation	39.78–46.51	1 (4%)	5.67	9 (36%)
Body weight (g) at preputial separation	188.28–256.19	2 (8%)	7.57	22 (88%)
Initial body weight (g)	45.47–59.81	9 (36%)	10.25	2 (8%)
Final body weight (g)	259.24–332.06	0 (0%)	7.47	14 (56%)
SVCG w/fluid (g)	0.295–0.719	3 (12%)	21.06	15 (60%)
Ventral prostate (g)	0.160–0.332	0 (0%)	22.32	5 (20%)
LABC (g)	0.447–0.855	2 (8%)	27.10	2 (8%)
Epididymis (g)	0.364–0.528	4 (16%)	16.39	2 (8%)
Liver (g)	9.99–15.35	0 (0%)	14.93	4 (16%)
Kidneys (g)	2.42–3.05	19 (76%)	14.76	1 (4%)
Pituitary (mg)	7.81–12.90	0 (0%)	15.98	7 (28%)
Adrenals (mg)	31.84–61.11	0 (0%)	22.77	1 (4%)
Thyroid (mg)	14.00–26.00	11 (44%)	23.63	3 (12%)
Serum T4 (ug/dL)	4.06–7.38	3 (12%)	27.46	1 (4%)
Serum testosterone (ng/mL)	0.260–3.960	0 (0%)	89.70	1 (4%)
Serum TSH (ng/mL)	4.21–24.11	9 (36%)	58.29	7 (28%)

Note: Performance criteria based on control substances only. Refer to OCSPP Test Guideline 890.1500 for a full listing of performance requirements. *Abbreviations:* CV = coefficient of variation; PND = post-natal day; SVCG = seminal vesicles + coagulating gland; LABC = levator ani/bulbocavernosus muscle; T4 = thyroxine; TSH = thyroid stimulating hormone. *n* = 25 studies.

Table 8
Overall *F*-tests (One-Way Analysis of Variance) of the effect of laboratory on endpoint variance. Results reported for endpoints with at least 25% of studies in violation of the maximum CV.

Endpoint	Numerator DF	Denominator DF	F-value	PR > *F*
Male Pubertal Assay				
Age (PND) at preputial separation	4	20	2.05	0.1253
Body weight (g) at preputial separation	4	20	0.68	0.6126
Final body weight (g)	4	20	2.83	0.0520
SVCG w/fluid (g)	4	20	1.34	0.2904
Epididymis, left (g)	4	20	1.27	0.3156
Epididymis, right (g)	4	20	1.67	0.1966
Pituitary (mg)	4	20	1.26	0.3182
Serum TSH (ng/mL)	4	20	0.96	0.4502
Female Pubertal Assay				
Age (PND) at vaginal opening	4	20	0.27	0.8928
Liver (g)	4	20	1.48	0.2455
Fish Short-Term Reproductive Assay				
Fertilization success (%)	2	19	0.06	0.9405

Abbreviations: DF = degrees of freedom; PND = postnatal day; SVCG = seminal vesicles + coagulating gland; TSH = thyroid stimulating hormone.

differences between laboratories at the $\alpha = 0.05$ level, suggesting that the maximum allowable CVs defined in the test guidelines warrant further examination.

3.2.2. Female Pubertal Assay

The Female Pubertal Assay utilizes a vehicle control as a comparator to the test chemical. As seen with the Male Pubertal Assay, performance criteria have been established for a number of pubertal development and growth endpoints. The number and percentage of studies failing these criteria are presented in Table 9. Summary statistics are provided in Tables A13 and A14. A majority (19 out of 25; 76%) of studies had at least 1 mean-range violation and 16 (64%) had at least 1 CV violation. In addition, 1 (4%) study had at least 3 mean-range violations and 2 (8%) had at least 3 CV violations. The overall mean of *Adrenals (mg)*, reported as 36.07 mg in Table A13, was outside of the acceptable mean range of 38.34–48.84 mg (Table 9). Three (12%) studies were outside the acceptable mean range for *Age (PND) at Vaginal Opening*, while 6 (24%) failed to meet the criterion for maximum CV. Model (1) was fitted to endpoints with at least 25% of studies in violation of the maximum CV. Two endpoints (Table 8) met this criterion. Based on the overall *F*-tests, there were no statistically-significant differences between laboratories at the $\alpha = 0.05$ level, suggesting, similar to the Male Pubertal Assay, that the maximum allowable CVs defined in the test guidelines warrant further examination.

Table 9
Performance criteria in the Female Pubertal Assay.

Endpoint	Acceptable mean range	Number (%) of studies failed	Maximum CV (%)	Number (%) of studies failed
Age (PND) at vaginal opening	30.67–35.62	3 (12%)	6.52	6 (24%)
Body weight (g) at vaginal opening	101.71–131.44	1 (4%)	13.94	1 (4%)
Final body weight (g)	104.86–204.55	0 (0%)	8.93	5 (20%)
Liver (g)	4.32–11.78	0 (0%)	13.13	8 (32%)
Kidneys (g)	0.95–2.20	0 (0%)	10.76	4 (16%)
Pituitary (mg)	5.86–12.08	0 (0%)	26.97	0 (0%)
Adrenals (mg)	38.34–48.84	19 (76%)	22.97	1 (4%)
Ovaries (mg)	36.54–114.77	0 (0%)	23.20	3 (12%)
Uterus, blotted (mg)	187.40–410.38	0 (0%)	37.73	0 (0%)
Thyroid (mg)	6.20–22.20	0 (0%)	38.58	0 (0%)
Serum T4 (ug/dL)	2.69–5.38	0 (0%)	29.39	1 (4%)

Note: Performance criteria based on control substances only. Refer to OCSPP Test Guideline 890.1450 for a full listing of performance requirements. *Abbreviations:* CV = coefficient of variation; PND = postnatal day; TSH = thyroid stimulating hormone. *n* = 25 studies.

3.2.3. Amphibian Metamorphosis Assay

Developmental and growth endpoints in the Amphibian Metamorphosis Assay are summarized in Tables A15 and A16. The number and percentage of studies failing the performance criteria defined in the test guidelines are presented in Table 10. Of the 25 studies that reported information on tadpole survival, 1 (4%) failed the performance criterion for control mortality. Of the 33 studies reporting developmental data, none failed the performance criterion for median Nieuwkoop and Faber (1994; NF) Stage on day 21 and 1 (3.03%) failed the criterion for NF Stage distribution at test termination.

3.2.4. Fish Short-Term Reproductive Assay

Growth endpoints and a summary of survival and reproduction in the Fish-Short Term Reproductive Assay are presented in Tables A17, A18, and A19. The number and percentage of studies failing the performance criteria defined in the test guidelines are presented in Table 11. Twenty-two studies reported information on survival and reproduction; 1 study (4.55%) failed the performance criterion for control mortality; 1 study (4.55%) failed the performance criterion for egg production, and 7 (31.82%) studies failed the performance criterion for fertilization success. At test termination, mean values for *Female Body Weight (g)* and *Male Body Weight (g)* ranged from 1.05–1.92 to 2.00–4.74, respectively. Also at test termination, mean values for *Female Body Length (mm)* and *Male Body Length (mm)* ranged from 1.05–1.92 to 2.00–4.74, respectively. Model (1) was fitted to endpoints with at least 25% of studies in violation of the maximum CV. Two endpoints (Table 8) met this criterion. Based on the overall *F*-test, there was no statistically-significant difference between laboratories at the $\alpha = 0.05$ level.

3.2.5. Uterotrophic Assay

Developmental and growth endpoints in the Uterotrophic Assay are summarized in Tables A20 and A21. The number and percentage of studies failing the performance criteria defined in the test guidelines are presented in Table 12. Three alternative assay designs were utilized by respondents providing access to control data: the immature model using oral gavage as the route of administration; the ovariectomized model using subcutaneous injection;

Table 10
Performance criteria in the Amphibian Metamorphosis Assay.

Endpoint	Acceptable values	No. of studies	No. (%) failed
Control mortality	⩽2 Tadpoles/replicate	25	1 (4.00%)
NF stage, day 21	Median NF stage ⩾ 57	33	0 (0.00%)
NF stage, day 21	90th–10th percentiles < 4 stages	33	1 (3.03%)

Note: Performance criteria based on control substances only. Refer to OCSPP Test Guideline 890.1100 for a full listing of performance requirements. *Abbreviations:* NF = Nieuwkoop and Faber (1994; NF) developmental stage.

Table 11
Performance criteria in the Fish Short-Term Reproductive Assay.

Endpoint	Acceptable values	No. of studies	No. (%) failed
Survival (%)	⩾90%	22	1 (4.55%)
Eggs/female/day	⩾15 Eggs/female/day in all replicates	22	1 (4.55%)
Fertilization success (%)	⩾95%	22	7 (31.82%)

Note: Performance criteria based on control substances only. Refer to OCSPP Test Guideline 890.1350 for a full listing of performance requirements.

and the ovariectomized model using oral gavage. All of the models are acceptable per the assay guidelines (OPPTS 890.1600). The only performance criterion to be met for acceptability of the assays is that the blotted uterine weight be less than 0.09% of body weight for the immature model or less than 0.04% of body weight for the ovariectomized model. As shown in Table 12, all studies met the performance criterion for blotted uterine weights for both animal models (immature vs. ovariectomized) and routes of administration (oral vs. subcutaneously).

3.2.6. Hershberger Assay

To meet the performance criteria for the Hershberger Assay, no more than 3 of the 10 tissue weights may exceed the CV in the control group (agonist assay-vehicle control, antagoinst assay-vehicle control + TP) and high dose tested. If 4 or more tissues exceed the CV, the assay is to be repeated. The performance criteria established for all 5 androgen dependent tissues and the number and percentage of studies failing these criteria are presented in Table 13. Developmental and growth endpoints are summarized in Tables A22 and A23. Of the 21 studies that reported information on the reference androgen agonist, 2 (9.5%) failed to meet the performance criteria for *Seminal Vesicles (mg)* and *Glans Penis (mg)*, 1 failed to meet the criteria for *Levator Ani/Bulbocavernosus Complex (LABC) (mg)* and *Cowpers Glands (mg)*, and 4 studies (19%) failed the criterion for *Ventral Prostate (mg)*. For anti-androgenic activity (vehicle control with 0.4 mg/kg TP), 20 studies were reported and 2 (10%) failed to meet the performance criterion for *Ventral Prostate (mg)*. Due to the limitations of the data presented in this manuscript, we could not evaluate whether the high dose levels of the test chemicals from the submitted studies also exceeded the specified CV, causing the studies to be repeated.

4. Discussion

The EDSP Tier 1 Endocrine Screening Battery is intended to be a suite of *in vitro* and *in vivo* assays to identify the potential of a chemical to interact with the estrogen, androgen, thyroid, or steroidogenesis systems. The initial set of List 1 chemicals provides enough data to conduct a detailed, rigorous, and comprehensive evaluation of the performance of the battery. Due to animal welfare concerns and the significant cost of running the screens (as currently designed, $750,000–$1,000,000 per chemical [US EPA, 2008]), the SAB/SAP recommendation to stop and review all testing methods is clearly sensible. What is unclear, however, is whether EPA will have the opportunity to conduct this review prior to issuance of List 2 test orders, as a 2009 House Appropriations Committee report (H.R. 2996. H. Rept. 111–180) directed the Agency to release a second list of no less than 100 chemicals for ESB testing, which EPA published in November 2010 (Federal Register, 2010).

Table 12
Performance criteria in the Uterotrophic Assay.

Endpoint	Performance criterion	No. of studies	No. (%) failed
Blotted uterine weights			
Immature model (oral gavage)	<0.09% of body weight	9	0 (0%)
Ovariectomized model (subcutaneous)	<0.04% of body weight	3	0 (0%)
Ovariectomized model (oral gavage)	<0.04% of body weight	14	0 (0%)

Note: Performance criteria based on control substances only. Refer to OCSPP Test Guideline 890.1600 for a full listing of performance requirements.

Table 13
Performance criteria in the Hershberger Assay.

Endpoint	Maximum CV (%)	No. of studies	No. (%) failed[a]
Androgenic (vehicle control)			
Seminal vesicles (mg)	40	21	2 (9.52%)
Ventral prostate (mg)	45	21	4 (19.05%)
LABC (mg)	30	21	1 (4.76%)
Cowpers glands (mg)	55	21	1 (4.76%)
Glans penis (mg)	22	21	2 (9.52%)
Anti-androgenic (vehicle control + 0.4 mg/kg TP)			
Seminal vesicles (mg)	40	20	0 (0.00%)
Ventral prostate (mg)	40	20	2 (10.00%)
LABC (mg)	20	20	1 (5.00%)
Cowpers glands (mg)	35	20	0 (0.00%)
Glans penis (mg)	17	20	1 (5.00%)
Anti-androgenic (vehicle control + 0.2 mg/kg TP)			
Seminal vesicles (mg)	40	1	0 (0.00%)
Ventral prostate (mg)	40	1	0 (0.00%)
LABC (mg)	20	1	0 (0.00%)
Cowpers glands (mg)	35	1	0 (0.00%)
Glans penis (mg)	17	1	0 (0.00%)

Note: Performance criteria based on control substances only. Refer to OCSPP Test Guideline 890.1400 for a full listing of performance requirements. *Abbreviations:* CV = coefficient of variation; LABC = levator ani/bulbocavernosus muscle; TP = testosterone propionate.

[a] The test guidelines require no more than 3 tissues from both the vehicle control and the high dose (data not provided) to exceed the maximum CV.

The analysis of List 1 control data presented here, along with concerns raised elsewhere (e.g., Borgert et al., 2011a,b), support the notion that EPA should be given the opportunity to review all testing methods before issuing test orders for the second list of chemicals. Given the frequency with which certain performance criteria were violated, one focus of that review should consider adjustments to the tolerance limits for fitted curve parameters, acceptable mean ranges and maximum CVs to better reflect biological variability. These performance criteria are intended to serve as a benchmark for determining the quality of the data submitted to EPA and also provide a way to judge the variability in each endpoint. Poorly-defined criteria will result in discarded data and wasted resources.

The Male Pubertal Assay highlights this concern clearly. During the assay validation program (EPA, 2007a), all participating laboratories had CVs for certain endpoints that exceeded the maximum allowed under the test guidelines. One laboratory failed 6 of 17 criteria, two laboratories failed 5, and another laboratory failed 4. Nonetheless, the criteria were not modified prior to List 1 test orders being issued. These inter-laboratory validation results are similar to the analysis presented here of a subset of the List 1 control data. All (25 out of 25) studies had at least 1 CV violation, 21 (84%) had at least 3 CV violations, and 8 (32%) had at least 5 CV violations. In addition, five endpoints (*Body Weight (g) at Preputial Separation*, *Final Body Weight (g)*, *SVCG w/Fluid (g)*, *Epididymis, left (g)*, and *Epididymis, right (g)*) had CVs averaged across studies that exceeded the maximum allowed by the test guidelines. The model fitting exercise demonstrated that the magnitudes of endpoint variances were relatively consistent across laboratories. Performance violations were therefore systemic (observable across laboratories, including those involved in the initial validation effort), not driven by one or two laboratories, and thus, very unlikely to have been caused by poor execution.

Acceptable mean ranges were also problematic in the Male Pubertal Assay. All studies had at least 1 mean-range violation, 9 (36%) studies had at least 3 mean-range violations, and 1 (4%) study had at least 5 mean-range violations. In addition, mean ranges reported in Table A12 tended to be wider than those observed in the EPA inter-laboratory validation study. For example, EPA reported a mean range of 39.50–43.92 days for *Age (PND) at Preputial Separation* (EPA, 2007a). In the present data compilation representing 25 assays, these mean values ranged from 43.18 to 48.00 days. Similarly, *Body Weight (g) at Preputial Separation* ranged from 198.57 to 236.49 g in the EPA inter-laboratory validation study, whereas these mean values ranged from 195.76 to 267.68 g in this analysis.

There are a number of factors which might contribute to the disparity between endpoint variances observed in this study and in the EPA inter-laboratory validation. These same factors might also explain why many of the acceptable mean ranges are too narrow and why most of the maximum CVs are too small. EPA used historical control data to set the performance criteria, which are "intended to cover 95% of the values likely to be encountered from acceptable laboratories" (EPA, 2007a). It is unclear, however, how much information was actually available for defining these targets. Based on comparisons between the current analysis ($\sim$25 studies) and EPA validation (4 studies), it is likely that the historical control data provided an incomplete picture of natural variability. Furthermore, permissible differences in study design and execution are certain to contribute to inter- and intra-laboratory variability. These differences include diet, housing, water source, dosing vehicle, species and strain of rat, and daily gavage administration, among others. Finally, one of the primary endpoints in the Male Pubertal Assay is *Age (PND) at Preputial Separation* and, since attainment of puberty is a subjective evaluation, inter- and intra-laboratory variability is to be expected (Stump et al., 2014).

The Female Pubertal Assay also illustrates the problem of poorly-defined performance criteria. During the assay validation program (EPA, 2007b), all participating laboratories had CVs for certain endpoints that exceeded the maximum allowed under the test guidelines. This inter-laboratory validation result is similar to the analysis presented here. A majority (16 out of 25; 64%) of studies had at least 1 CV violation and 2 (8%) had at least 3 CV violations. As seen in the Male Pubertal Assay, there was insufficient evidence to conclude that endpoint variances differed significantly between laboratories. Acceptable mean ranges were also problematic: 19 out of 25 (76%) studies had at least 1 mean-range violation and 1 (4%) study had at least 3 mean-range violations. Mean ranges reported in Table A14 were consistently wider than those observed in the EPA inter-laboratory validation study. EPA reported a mean range of 31.56–34.21 days for *Age (PND) at Vaginal Opening*. Mean values for this same endpoint ranged from 32.73 to 36.79 days in the present study. Similarly, *Body Weight (g) at Vaginal Opening* ranged from 109.67 to 119.84 g in the EPA inter-laboratory validation study, whereas these mean values ranged from 105.53 to 137.08 g in this analysis. Disparities were also seen in *Kidneys (g)* (mean range 1.15–1.49 vs. 1.47–1.67), *Adrenals (mg)* (mean range 31.60–43.56 vs. 41.00–49.00), and *Serum T4 (ug/dL)* (mean range 2.97–5.04 vs. 4.77–7.94).

For the Female Pubertal Assay, the same factors mentioned above likely contribute to the inconsistency between endpoint variances observed in this study and in the EPA inter-laboratory validation, as well as explain why many of the maximum CVs are too small. The same permissible differences in study design and execution also contribute to inter- and intra-laboratory variability. The subjective evaluation of attainment of puberty (*Age (PND) at Vaginal Opening* and *Body Weight (g) at Vaginal Opening*) is also a source of variation.

Performance criteria were less problematic in the Amphibian Metamorphosis Assay. Of the 25 studies that reported information on tadpole survival, 1 (4%) failed the performance criterion for control mortality. Of the 33 studies reporting developmental data, none failed the performance criterion for median NF Stage on day 21 and 1 (3.03%) failed the criterion for NF Stage distribution at test termination. Performance criteria were reasonably well-defined in the Fish Short-Term Reproductive Assay. Twenty-two studies reported information on survival and reproduction; 1 study (4.55%) failed the performance criterion for control mortality; 1 study (4.55%) failed the performance criterion for egg production, and 7 (31.82%) studies failed the performance criterion for fertilization success.

Performance criteria for the Uterotrophic Assay were met in all studies for both immature and ovariectomized models, and both routes of administration. The extensive validation of the assay (Kanno et at., 2001, 2003a,b; OECD, 2001; Owens et al., 2003) likely contributes to the accurate performance criteria contained in the guidelines. Performance criteria were reasonably well-defined in the Hershberger Assay. However, the maximum allowable CV for *Ventral Prostate (mg)* was problematic, particularly for the agonist assay. In addition, data were only available from a single anti-androgenic study utilizing the 0.2 mg/kg/day dose of testosterone propionate. As such, results reported here should not be viewed as a rigorous evaluation of the suitability of these criteria. Lastly, performance criteria based on the CV must also be met by the high-dose treatment group. Since this analysis was restricted to control data, concerns remain about the variability in androgen-dependent tissues exposed to treated doses being greater than that found in the controls. This is a distinct possibility and should be considered in a comprehensive evaluation of the suitability of performance criteria.

Tolerance limits for fitted curve parameters and hormone concentrations were problematic in many of the *in vitro* screens. For instance, 31% of runs reported estimates of *Hill Slope* $(log_{10}(M)^{-1})$ for the weak positive control outside the tolerance limits in the ER Binding Assay. An expansion of the lower limit from -1.1 to -1.3 would encompass all of the reported findings. A majority of runs (62% and 67%) failed both performance criteria for the very weak agonist (17α-methyltestosterone) in the ERTA Assay. In addition, the ranges of acceptable values for *Hill Slope* $(log_{10}(M)^{-1})$ were problematic for the strong agonist (17β-estradiol) and weak agonist (17α-estradiol). In the Aromatase Assay, 25% of runs reported unacceptable estimates of $log(IC_{50})$ for the positive control; there was approximately an equal number of values above and below the acceptable range. A slight modification of the lower limit from -7.3 to -7.4 would capture more than half of the out-of-range estimates. In the Steroidogenesis Assay, 41% of runs reported unacceptable levels of *Estradiol* production for the positive control inhibitor (1 μM Prochloraz) and 25% of runs reported basal production of *Estradiol* below the acceptable minimum.

The discussion devoted to performance criteria highlights an important question that has not been addressed by EPA: when does a violation(s) render a study unreliable? Although many of the violations are minor (i.e., narrowly missing the acceptable mean range or maximum CV), little to no guidance exists on the degree to which assays need to be re-started for failing to meet performance criteria. We recommend that a second focus of the suggested EPA review of the EDSP ESB be to provide detailed, assay-specific direction on when results should be discarded. This point is in agreement with a FIFRA Scientific Advisory Panel (SAP), which suggested that specific guidance be developed to indicate the point at which performance violations render the data unusable (FIFRA SAP, 2013). From our experience responding to List 1 test orders, we are aware that many assays were re-started due to concerns over missed performance criteria. Although the overall incidence of assay re-starts is unknown, this undoubtedly increased animal use and overall costs of the EDSP program and contributed to delays. However, those expenditures were not included in the estimated costs of the eleven EDSP ESB assays (US EPA, 2009b). In order to more accurately estimate the costs of the Tier 1 battery and to understand how best to refine guidance for re-starting assays, we encourage EPA to consult the laboratories that conducted List 1 screening to determine the incidence that assays were re-started and the associated costs in terms of animals, resources, and time to complete the screening.

The FIFRA SAP also highlighted the frequency with which certain performance criteria were violated and discussed permissible differences in study design and execution that are likely to contribute to inter- and intra-laboratory variability (FIFRA SAP, 2013). Consistent with this dialogue, we recommend that a third focus of the suggested EPA review of the EDSP ESB should be to identify protocol differences with a large influence on endpoint variance. After identification of these influential factors, experimental guidelines could be re-defined such that endpoint variances are reduced and performance criteria are violated less frequently.

A phased implementation of the EDSP should serve, over time, to improve the specificity, sensitivity, and interpretability of the ESB battery. At this time, EPA should review all List 1 control data submitted to the Agency (52 chemicals). Updating the performance criteria and protocols based on such a review would eliminate many of the violations that occurred from happening again with the second list of chemicals. This, however, will have little or no benefit if EPA is not allowed to follow the SAB/SAP recommendation to complete the review prior to issuing test orders for the second list of chemicals. Indeed, if this decision is made, all benefits of this and similar efforts will go unrealized, including some benefits of EPA's long-term effort to develop high-throughput methods for use in the EDSP, an effort now known as "EDSP-21" (U.S. EPA, 2011b).

The EDSP-21 program is currently most well developed for the estrogenic pathway, with a stated goal of eventually replacing the Uterotrophic Assay in the Tier 1 battery in order to reduce animal use and costs and increase the overall efficiency of the program. However, the analyses presented here reveal that this particular substitution is likely to affect the EDSP program opposite of what was intended, for the following reason. Justifiably, EPA and others agree that in principle, *in vivo* assays should be given more weight than *in vitro* assays in the interpretation of the Tier 1 screening results (Borgert et al., 2011a; US EPA, 2011a). Hence, replacing an *in vivo* with *in vitro* assays would shift interpretational emphasis to the other *in vivo* endpoints remaining in the battery. Because the Tier 1 *in vivo* screens vary widely in meeting performance criteria, as shown here, this particular substitution would shift interpretational emphasis from the most reliable *in vivo* assay in the battery (Uterotrophic) to the least reliable (Pubertal Assays). Unless EPA is given the time and resources to fully evaluate performance of the battery, it will be difficult to make improvements to the battery without incurring unforeseen liabilities.

Conflict of Interest

Some substances undergoing EDSP screening are compounds produced by Monsanto Company, The Dow Chemical Company, and Bayer CropScience. However, this manuscript applies to all substances impacted by EPA EDSP test orders, not merely those produced by the authors' employers. This manuscript has been reviewed in accordance with the peer- and administrative-review policies of the authors' organizations. The views expressed here are those of the authors and do not necessarily reflect the opinions and/or policies of their employers. There are no contractual relations or proprietary considerations that restrict dissemination of the research findings of the authors. C.J. Borgert and L.D. Stuchal are independent scientists/consultants who received financial support for portions of this project from the Endocrine Policy Forum. Time spent by other co-authors was supported by their respective employers or a personal contribution.

Transparency Document

The Transparency document associated with this article can be found in the online version.

Acknowledgments

These analyses would not have been possible without the considerable contribution of control data from Tier 1 assays provided by:

Simon Dinehart, Ph.D., Stephen Hicks – ABC Laboratories, Inc.

Suzanne Schneider, Ph.D., Jeremy Claude, Bridgett Page, Susan Palmer, Lene Petersen, Ph.D. – Wildlife International Aquatics Department.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.yrtph.2015.05.028.

References

Borgert, C.J., Mihaich, E.M., Ortego, L.S., Bentley, K.S., Homes, C.M., Levine, S.L., Becker, R.A., 2011a. Hypothesis-driven weight of evidence framework for evaluating data within the US EPA's Endocrine Disruptor Screening Program. Regul. Toxicol. Pharmacol. 61, 185–191.

Borgert, C.J., Mihaich, E.M., Quill, T.F., Marty, M.S., Levine, S.L., Becker, R.A., 2011b. Evaluation of EPA's Tier 1 Endocrine Screening Battery and recommendations for improving the interpretation of screening results. Regul. Toxicol. Pharmacol. 59, 397–411.

Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC), 1998. Final Report, 1998. http://www.epa.gov/scipoly/oscpendo/pubs/edspoverview/finalrpt.htm.

Federal Register, 2010. 75 US Federal Register. 70248–70254, EPA, Notice, Endocrine Disruptor Screening Program; Second List of Chemicals, November 17, 2010.

FIFRA SAP, 2013. Transmittal of meeting minutes from the FIFRA Scientific Advisory Panel held May 21–23, 2013 on "Endocrine Disruptor Screening Program (EDSP) Tier 1 Screening Assays and Battery Performance". Memorandum to the Office of Science Coordination and Policy and the Office of Pesticide Programs, US EPA, data August 21, 2013. http://www.epa.gov/scipoly/sap/meetings/2013/may/052113minutes.pdf.

Hill, A.V., 1910. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. J. Physiol. (Lond.) 40, 4–7.

Kanno, J., Onyon, L., Haseman, J., Fenner-Crisp, P., Ashby, J., Owens, W., 2001. The OECD program to validate the rat uterotrophic bioassay to screen compounds for in vivo estrogenic responses: phase 1. Environ. Health Perspect. 109, 785–794.

Kanno, J., Onyon, L., Peddada, S., Ashby, J., Jacob, E., Owens, W., 2003a. The OECD program to validate the rat uterotrophic bioassay to screen compounds for in vivo estrogenic responses: phase 2. Dose response studies. Environ. Health Perspect. 111 (12), 1550–1558.

Kanno, J., Onyon, L., Peddada, S., Ashby, J., Jacob, E., Owens, W., 2003b. The OECD program to validate the rat uterotrophic bioassay to screen compounds for in vivo estrogenic responses: phase 2. Single coded dose studies. Environ. Health Perspect. 111 (12), 1530–1549.

Marty, M.S., Carney, E.W., Rowlands, J.C., 2011. Endocrine disruption: historical perspectives and its impact on the future of toxicology testing. Toxicol. Sci. 120 (Suppl. 1), S93–108.

Nieuwkoop, P.D., Faber, J., 1994. Normal Table of *Xenopus laevis* (daudin). Garland Press, New York, USA.

OECD (Organisation for Economic Co-operation and Development), 2001. Progress of work on the validation of the rodent uterotrophic bioassay: phase two status report (including statistical evaluation of the results). ENV/JM/TG/EDTA (2001) 2, April, 2001. Endocrine Disruptor Testing and Assessment. OECD, Paris. 45 p.

Owens, W., Ashby, J., Odum, J., Onyon, L., 2003. The OECD program to validate the rat uterotrophic bioassay. Phase 2: dietary phytoestrogen analyses. Environ. Health Perspect. 111, 1559–1567.

SAS, 2012. Software Release 9.4 (TS1M1). Copyright 2002–2012 by SAS Institute Inc., Cary, North Carolina.

Stump, D.G., O'Connor, J.C., Lewis, J.M., Marty, S.M., 2014. Key lessons from performance of the U.S. EPA Endocrine Discruptor Screening Program (EDSP) Tier 1 Male and Female Pubertal Assays.

The Keystone Center, 1996. Keystone Convening Report Regarding the Formation of the Endocrine Disruptor Testing and Advisory Committee. http://www.epa.gov/scipoly/oscpendo/pubs/edsparchive/keystone.htm.

US EPA, 1996. Food Quality Protection Act. Public Law 104–170, 104th Congress. http://www.epa.gov/pesticides/regulating/laws/fqpa/.

US EPA, 1999. Review of EPA's Proposed Environmental Endocrine Disruptor Screening Program; Review of the Endocrine Disruptor Screening Program by a Joint Subcommittee of the Science Advisory Board and Scientific Advisory Panel. EPA-SAB-EC-99-013. http://www.epa.gov/endo/pubs/sab_sap_report.pdf.

US EPA, 2007. Integrated summary report for validation of a test method for assessment of pubertal development and thyroid function in juvenile male rats as a potential screen in the Endocrine Disruptor Screening Program Tier-1 Battery. http://www.epa.gov/endo/pubs/male_pubertal_isr.pdf.

US EPA, 2007. Integrated summary report for validation of a test method for assessment of pubertal development and thyroid function in juvenile female rats as a potential screen in the Endocrine Disruptor Screening Program Tier-1 Battery. http://www.epa.gov/endo/pubs/female_isr_v4.1c.pdf.

US EPA, 2008. Comments of the Chemical Producers and Distributors Association et al., on EPA's Information Collection Request. Submitted May 22, 2008. Docket: EPA-HQ-OPPT-2007-1081-0020.

US EPA, 2009. Endocrine Disruptor Screening Program: Tier 1 Screening Order Issuing Announcement. 74 Fed. Reg. 54422, Oct. 2, 2009.

US EPA, 2009b. Supporting Statement for an Information Request [EPA ICR No. 2249.01, OMB Control No. 2070–new]; Docket No. EPA-HQ-2007-1081-0017. U.S. Environmental Protection Agency, Washington, D.C., 19 p.

US EPA, 2011. Endocrine Disruptor Screening Program. Weight-of-Evidence: Evaluating Results of EDSP Tier 1 Screening to Identify the Need for Tier 2 Testing. http://www.regulations.gov/#!documentDetail;D=EPA-HQ-OPPT-2010-0877-0021.

US EPA, 2011. Endocrine Disruptor Screening Program for the 21st century: (EDSP21 Work Plan) the incorporation of in silico models and in vitro high throughput assays in the Endocrine Disruptor Screening Program (EDSP) for Prioritization and Screening. Office of Chemical Safety and Pollution Prevention, US Environmental Protection Agency, Washington DC, 20460. http://www.epa.gov/endo/pubs/edsp21_work_plan_summary%20_overview_final.pdf.